

LNCS 3332

Kiyoharu Aizawa
Yuichi Nakamura
Shin'ichi Satoh (Eds.)

Advances in Multimedia Information Processing – PCM 2004

**5th Pacific Rim Conference on Multimedia
Tokyo, Japan, November/December 2004
Proceedings, Part II**

2 Part II

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

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

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

New York University, NY, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Kiyoharu Aizawa Yuichi Nakamura
Shin'ichi Satoh (Eds.)

Advances in Multimedia Information Processing – PCM 2004

5th Pacific Rim Conference on Multimedia
Tokyo, Japan, November 30 – December 3, 2004
Proceedings, Part II

Volume Editors

Kiyoharu Aizawa

Department of Frontier Informatics, The University of Tokyo

5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

E-mail: aizawa@hal.t.u-tokyo.ac.jp

Yuichi Nakamura

Academic Center for Computation and Media Studies, Kyoto University

Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

E-mail: yuichi@media.kyoto-u.ac.jp

Shin'ichi Satoh

National Institute of Informatics

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

E-mail: satoh@nii.ac.jp

Library of Congress Control Number: 2004115461

CR Subject Classification (1998): H.5.1, H.3, H.5, C.2, H.4, I.3, K.6, I.7, I.4

ISSN 0302-9743

ISBN 3-540-23977-4 Springer Berlin Heidelberg New York

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

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004

Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH

Printed on acid-free paper SPIN: 11363255 06/3142 5 4 3 2 1 0

Preface

Welcome to the proceedings of the 5th Pacific Rim Conference on Multimedia (PCM 2004) held in Tokyo Waterfront City, Japan, November 30–December 3, 2004. Following the success of the preceding conferences, PCM 2000 in Sydney, PCM 2001 in Beijing, PCM 2002 in Hsinchu, and PCM 2003 in Singapore, the fifth PCM brought together the researchers, developers, practitioners, and educators in the field of multimedia. Theoretical breakthroughs and practical systems were presented at this conference, thanks to the support of the IEEE Circuits and Systems Society, IEEE Region 10 and IEEE Japan Council, ACM SIGMM, IEICE and ITE.

PCM 2004 featured a comprehensive program including keynote talks, regular paper presentations, posters, demos, and special sessions. We received 385 papers and the number of submissions was the largest among recent PCMs. Among such a large number of submissions, we accepted only 94 oral presentations and 176 poster presentations. Seven special sessions were also organized by world-leading researchers. We kindly acknowledge the great support provided in the reviewing of submissions by the program committee members, as well as the additional reviewers who generously gave their time. The many useful comments provided by the reviewing process must have been very valuable for the authors' work.

This conference would never have happened without the help of many people. We greatly appreciate the support of our strong organizing committee chairs and advisory chairs. Among the chairs, special thanks go to Dr. Ichiro Ide and Dr. Takeshi Naemura who smoothly handled publication of the proceedings with Springer. Dr. Kazuya Kodama did a fabulous job as our Web master.

September 2004

Kiyoharu Aizawa
Yuichi Nakamura
Shin'ichi Satoh
Masao Sakauchi

PCM 2004 Organization

Organizing Committee

Conference Chair	Masao Sakauchi <i>NII/The Univ. of Tokyo</i>
Program Co-chairs	Kiyoharu Aizawa <i>The Univ. of Tokyo</i> Yuichi Nakamura <i>Kyoto Univ.</i> Shin'ichi Satoh <i>NII</i>
Poster/Demo Co-chairs	Yoshinari Kameda <i>Univ. of Tsukuba</i> Takayuki Hamamoto <i>Tokyo Univ. of Science</i>
Financial Co-chairs	Nobuji Tetsutani <i>Tokyo Denki Univ.</i> Hirohisa Jozawa <i>NTT Resonant</i>
Publicity Co-chairs	Noboru Babaguchi <i>Osaka Univ.</i> Yoshiaki Shishikui <i>NHK</i>
Publication Co-chairs	Takeshi Naemura <i>The Univ. of Tokyo</i> Ichiro Ide <i>Nagoya Univ.</i>
Registration Chair	Ryoichi Kawada <i>KDDI</i>
Web Chair	Kazuya Kodama <i>NII</i>
USA Liaison	Tsuhan Chen <i>CMU</i>
Korea Liaison	Yo-Sung Ho <i>K-JIST</i>
Advisory Committee	Sun-Yuan Kung <i>Princeton Univ.</i> Hong-Jiang Zhang <i>Microsoft Research Asia</i> Masayuki Tanimoto <i>Nagoya Univ.</i>

Mark Liao
Academia Sinica
Hiroschi Harashima
The Univ. of Tokyo

Program Committee

Masao Aizu

Canon

Laurent Amsaleg

IRISA-CNRS

Yasuo Ariki

Kobe Univ.

Alberto Del Bimbo

Univ. of Florence

Nozha Boujemaa

INRIA Rocquencourt

Jihad F. Boulos

American Univ. of Beirut

Tat-Seng Chua

National Univ. of Singapore

Chabane Djeraba

LIFL

Toshiaki Fujii

Nagoya Univ.

Yihong Gong

NEC Laboratories America

Patrick Gros

IRISA-CNRS

William Grosky

Univ. of Michigan, Dearborn

Alexander G. Hauptmann

CMU

Yun He

Tsinghua Univ.

Xian-Sheng Hua

Microsoft Research Asia

Takashi Ida

Toshiba

Hiroyuki Imaizumi

NHK-ES

Takashi Itoh

Fujitsu

Alejandro Jaimes

FX Pal Japan, Fuji Xerox

Mohan S. Kankanhalli

National Univ. of Singapore

Norio Katayama

NII

Jiro Katto

Waseda Univ.

Asanobu Kitamoto

NII

Hitoshi Kiya

Tokyo Metropolitan Univ.

Byung-Uk Lee

Ewha Univ.

Sang-Wook Lee

Seoul National Univ.

Michael Lew

Univ. of Leiden

Mingjing Li

Microsoft Research Asia

Rainer Lienhart

Univ. Augsburg

Wei-Ying Ma

Microsoft Research Asia

Michihiko Minoh

Kyoto Univ.

Hiroschi Murase

Nagoya Univ.

Chong-Wah Ngo

City Univ. of Hong Kong

Satoshi Nogaki

NEC

Vincent Oria

New Jersey Institute of Technology

Rae-Hong Park

Sogang Univ.

Helmut Prendinger

NII

Jong-Beom Ra

KAIST

Takahiro Saito*Kanagawa Univ.***Philippe Salembier***Univ. Politecnica de Catalunya***Nicu Sebe***Univ. of Amsterdam***Timothy K. Shih***Tamkang Univ.***John Smith***IBM T.J. Watson Research Center***Kenji Sugiyama***Victor***Ming Ting Sun***Univ. of Washington***Seishi Takamura***NTT***Qi Tian***Institute for Infocomm Research***Luis Torres***Univ. Politecnica de Catalunya***Marcel Worring***Univ. of Amsterdam***Yoshihisa Yamada***Mitsubishi Electric***Naokazu Yokoya***Nara Institute of Science
and Technology***Additional Reviewers**

Frank Aldershoff

Hirofumi Aoki

Yukihiro Bandoh

Istvan Barakonyi

Stefano Berretti

Marco Bertini

Lei Chen

Keiichi Chono

He Dajun

Manolis Delakis

Takuya Funatomi

Guillaume Gravier

Keiji Gyohten

Reiko Hamada

Mei Han

Atsushi Hatabu

Ngoh Lek Heng

Xian-Sheng Hua

Lim Joo Hwee

Masaaki Iiyama

Mitsuo Ikeda

Kiyohiko Ishikawa

Hironori Ito

Yoshimichi Ito

Junko Itou

Wei Jiang

Wanjun Jin

Koh Kakusho

Masayuki Kanbara

Yutaka Kaneko

Ewa Kijak

Jonghwa Kim

Hideaki Kimata

Takahiro Kimoto

Koichi Kise

Masaki Kitahara

Takayuki Kitasaka

Zhiwei Li

Lie Lu

Yufei Ma

Keigo Majima

Takafumi Marutani

Yutaka Matsuo

Toshihiro Minami

Yoshihiro Miyamoto

Seiya Miyazaki

Kensaku Mori

Takeshi Mori

Satoshi Nishiguchi

Takayuki Onishi

Wei-Tsang Ooi

Jia-wei Rong

Tomasz M. Rutkowski

Shinichi Sakaida

Tomokazu Sato

Susumu Seki

Yuzo Senda

Fumihisa Shibata

Tomokazu Takahashi

Hung-Chuan Teh

Qiang Wang

Kaoru Watanabe

Joost van de Weijer

Jun Wu

Huaxin Xu

Keisuke Yagi

Itheri Yahiaoui

Kazumasa Yamazawa

Table of Contents, Part II

Volume II

Application of Video Browsing to Consumer Video Browsing Products

Generic Summarization Technology for Consumer Video	1
<i>Masaru Sugano, Yasuyuki Nakajima, Hiromasa Yanagihara, Akio Yoneyama</i>	
Movie-in-a-Minute: Automatically Generated Video Previews	9
<i>Mauro Barbieri, Nevenka Dimitrova, Lalitha Agnihotri</i>	
Automatic Sports Highlights Extraction with Content Augmentation	19
<i>Kongwah Wan, Jinjun Wang, Changsheng Xu, Qi Tian</i>	
Audio-Assisted Video Browsing for DVD Recorders	27
<i>Ajay Divakaran, Isao Otsuka, Regunathan Radhakrishnan, Kazuhiko Nakane, Masaharu Ogawa</i>	

Watermarking (I)

A Robust Image Watermarking Technique for JPEG Images Using QuadTrees	34
<i>Kil-Sang Yoo, Mi-Ae Kim, Won-Hyung Lee</i>	
A Fragile Watermarking Technique for Image Authentication Using Singular Value Decomposition	42
<i>Vivi Oktavia, Won-Hyung Lee</i>	
Visual Cryptography for Digital Watermarking in Still Images	50
<i>Gwo-Chin Tai, Long-Wen Chang</i>	
A New Object-Based Image Watermarking Robust to Geometrical Attacks	58
<i>Jung-Soo Lee, Whoi-Yul Kim</i>	
A Selective Image Encryption Scheme Based on JPEG2000 Codec	65
<i>Shiguo Lian, Jinsheng Sun, Dengfeng Zhang, Zhiquan Wang</i>	
VQ-Based Gray Watermark Hiding Scheme and Genetic Index Assignment	73
<i>Feng-Hsing Wang, Jeng-Shyang Pan, Lakhmi Jain, Hsiang-Cheh Huang</i>	

User Interface (I)

Advanced Paper Document in a Projection Display	81
<i>Kwangjin Hong, Keechul Jung</i>	
Improving Web Browsing on Small Devices Based on Table Classification	88
<i>Chong Wang, Xing Xie, Wenyan Wang, Wei-Ying Ma</i>	
A Java-Based Collaborative Authoring System for Multimedia Presentation	96
<i>Mee Young Sung, Do Hyung Lee</i>	
Object Tracking and Object Change Detection in Desktop Manipulation for Video-Based Interactive Manuals	104
<i>Yosuke Tsubuku, Yuichi Nakamura, Yuichi Ohta</i>	
Design of an Integrated Wearable Multimedia Interface for In-Vehicle Telematics	113
<i>Seongil Lee, Sang Hyuk Hong</i>	

Content-Based Image Retrieval

Video Scene Retrieval with Sign Sequence Matching Based on Audio Features	121
<i>Keisuke Morisawa, Naoko Nitta, Noboru Babaguchi</i>	
Architecture and Analysis of Color Structure Descriptor for Real-Time Video Indexing and Retrieval	130
<i>Jing-Ying Chang, Chung-Jr Lian, Hung-Chi Fang, Liang-Gee Chen</i>	
Automatic Salient-Object Extraction Using the Contrast Map and Salient Points	138
<i>SooYeong Kwak, ByoungChul Ko, Hyeran Byun</i>	
Shape-Based Image Retrieval Using Invariant Features	146
<i>Jong-Seung Park, TaeYong Kim</i>	
Visual Trigger Templates for Knowledge-Based Indexing	154
<i>Alejandro Jaimes, Qinhui Wang, Noriji Kato, Hitoshi Ikeda, Jun Miyazaki</i>	
Browsing and Similarity Search of Videos Based on Cluster Extraction from Graphs	162
<i>Seiji Hotta, Senya Kiyasu, Sueharu Miyahara</i>	
Correlation Learning Method Based on Image Internal Semantic Model for CBIR	172
<i>Lijuan Duan, Guojun Mao, Wen Gao</i>	

Controlling Concurrent Accesses in Multimedia Databases for Decision Support	180
<i>Woochun Jun, Suk-ki Hong</i>	
Image Retrieval by Categorization Using LVQ Network with Wavelet Domain Perceptual Features	188
<i>M.K. Bashar, Noboru Ohnishi, Kiyoshi Agusa</i>	
A Content-Based News Video Browsing and Retrieval System: NewsBR	197
<i>Huayong Liu, Zhang Hui</i>	
A News Video Browser Using Identical Video Segment Detection.....	205
<i>Fuminori Yamagishi, Shin'ichi Satoh, Masao Sakauchi</i>	
Pseudo Relevance Feedback Based on Iterative Probabilistic One-Class SVMs in Web Image Retrieval	213
<i>Jingrui He, Mingjing Li, Zhiwei Li, Hong-Jiang Zhang, Hanghang Tong, Changshui Zhang</i>	
Robust Video Similarity Retrieval Using Temporal MIMB Moments	221
<i>Duan-Yu Chen, Suh-Yin Lee</i>	
Complete Performance Graphs in Probabilistic Information Retrieval	229
<i>N. Sebe, D.P. Huijsmans, Q. Tian, T. Gevers</i>	
A New MPEG-7 Standard: Perceptual 3-D Shape Descriptor	238
<i>Duck Hoon Kim, In Kyu Park, Il Dong Yun, Sang Uk Lee</i>	
News Video Summarization Based on Spatial and Motion Feature Analysis	246
<i>Wen-Nung Lie, Chun-Ming Lai</i>	
Sports (II)	
SketchIt: Basketball Video Retrieval Using Ball Motion Similarity	256
<i>Sitaram Bhagavathy, Motaz El-Saban</i>	
Implanting Virtual Advertisement into Broadcast Soccer Video	264
<i>Changsheng Xu, Kong Wah Wan, Son Hai Bui, Qi Tian</i>	
Automatic Video Summarization of Sports Videos Using Metadata	272
<i>Yoshimasa Takahashi, Naoko Nitta, Noboru Babaguchi</i>	
Classification of Frames from Broadcasted Soccer Video and Applications	281
<i>Qing Tang, Jesse S. Jin, Haiping Sun</i>	

An Online Learning Framework for Sports Video View Classification 289
*Jun Wu, XianSheng Hua, JianMin Li, Bo Zhang,
 HongJiang Zhang*

A Semantic Description Scheme of Soccer Video Based on MPEG-7 298
Lin Liu, Xiuzi Ye, Min Yao, Sanyuan Zhang

Spatio-temporal Pattern Mining in Sports Video 306
Dong-Jun Lan, Yu-Fei Ma, Wei-Ying Ma, Hong-Jiang Zhang

Archiving Tennis Video Clips Based on Tactics Information 314
*Jenny R. Wang, Nandan Prameswaran, Xinguo Yu,
 Changsheng Xu, Qi Tian*

Network (III)

Performance Evaluation of High-Speed TCP Protocols with Pacing 322
Young-Soo Choi, Kang-Won Lee, You-Ze Cho

Time-Triggered and Message-Triggered Object Architecture
 for Distributed Real-Time Multimedia Services 330
Doo-Hyun Kim, Eun Hwan Jo, Moon Hae Kim

A Novel Multiple Time Scale Congestion Control Scheme
 for Multimedia Traffic 338
*Hao Yin, Chuang Lin, Ting Cui, Xiao-meng Huang,
 Zhang-xi Tan*

Dynamic Programming Based Adaptation of Multimedia Contents
 in UMA 347
Truong Cong Thang, Yong Ju Jung, Yong Man Ro

Performance Analysis of MAC Protocol for EPON Using OPNET 356
Min-Suk Jung, Jong-hoon Eom, Sung-Ho Kim

Adaptive FEC Control
 for Reliable High-Speed UDP-Based Media Transport 364
Young-Woo Kwon, Hyeyoung Chang, JongWon Kim

Efficient Overlay Network for P2P Content Sharing Based
 on Network Identifier 373
Chanmo Park, JongWon Kim

A Forward-Backward Voice Packet Loss Concealment Algorithm
 for Multimedia over IP Network Services 381
*Mi Suk Lee, Hong Kook Kim, Seung Ho Choi, Eung Don Lee,
 Do Young Kim*

A Delay-Based End-to-End Congestion Avoidance Scheme for Multimedia Networks	389
<i>Li Yan, Bin Qiu, Lichang Che</i>	

Dynamic Bandwidth Allocation for Internet Telephony	397
<i>Yiu-Wing Leung</i>	

Streaming (I)

A Broadcasting Technique for Holographic 3D Movie Using Network Streaming	405
<i>Kunihiko Takano, Koki Sato, Ryoji Wakabayashi, Kenji Muto, Kazuo Shimada</i>	

Coping with Unreliable Peers for Hybrid Peer-to-Peer Media Streaming	415
<i>Sung-Hoon Sohn, Yun-Cheol Baek, Juno Chang</i>	

Evaluation of Token Bucket Parameters for VBR MPEG Video	423
<i>Sang-Hyun Park, Yoon Kim</i>	

Perceptual Video Streaming by Adaptive Spatial-Temporal Scalability	431
<i>Wei Lai, Xiao-Dong Gu, Ren-Hua Wang, Li-Rong Dai, Hong-Jiang Zhang</i>	

A Proxy Caching System Based on Multimedia Streaming Service over the Internet	439
<i>Hui Guo, Jingli Zhou, Dong Zeng, Shengsheng Yu</i>	

Simulation and Development of Event-Driven Multimedia Session	447
<i>Nashwa Abdel-Baki, Hans Peter Großmann</i>	

Applying Linux High-Availability and Load Balancing Servers for Video-on-Demand (VOD) Systems	455
<i>Chao-Tung Yang, Ko-Tzu Wang, Kuan-Ching Li, Liang-Teh Lee</i>	

Visual Content Mining in Multimedia Documents

Indexing Issues in Supporting Similarity Searching	463
<i>Hanan Samet</i>	

Efficient Visual Content Retrieval and Mining in Videos	471
<i>Josef Sivic, Andrew Zisserman</i>	

Fast and Robust Short Video Clip Search for Copy Detection	479
<i>Junsong Yuan, Ling-Yu Duan, Qi Tian, Surendra Ranganath, Changsheng Xu</i>	

Mining Large-Scale Broadcast Video Archives
Towards Inter-video Structuring 489
Norio Katayama, Hiroshi Mo, Ichiro Ide, Shin'ichi Satoh

Sample Selection Strategies for Relevance Feedback
in Region-Based Image Retrieval 497
Marin Ferecatu, Michel Crucianu, Nozha Boujemaa

Compression (I)

Comparison of Two Different Approaches to Detect Perceptual Noise
for MPEG-4 AAC 505
Cheng-Hsun Yu, Shingchern D. You

Optimum End-to-End Distortion Estimation
for Error Resilient Video Coding 513
Yuan Zhang, Qingming Huang, Yan Lu, Wen Gao

Enhanced Stochastic Bit Reshuffling
for Fine Granular Scalable Video Coding 521
Wen-Hsiao Peng, Tihao Chiang, Hsueh-Ming Hang, Chen-Yi Lee

High-Performance Motion-JPEG2000 Encoder
Using Overlapped Block Transferring and Pipelined Processing 529
*Byeong-Doo Choi, Min-Cheol Hwang, Ju-Hun Nam,
Kyung-Hoon Lee, Sung-Jea Ko*

Low-Power Video Decoding
for Mobile Multimedia Applications 537
Seongsoo Lee, Min-Cheol Hong

Adaptive Macro Motion Vector Quantization 545
Luis A. da Silva Cruz

Face, Gesture, and Behavior (II)

Automatic, Effective, and Efficient 3D Face Reconstruction
from Arbitrary View Image 553
*Changhu Wang, Shuicheng Yan, Hua Li, Hongjiang Zhang,
Mingjing Li*

Recognition and Retrieval of Face Images
by Semi-supervised Learning 561
Kohei Inoue, Kiichi Urahama

3-D Facial Expression Recognition-Synthesis
on PDA Incorporating Emotional Timing 569
*Doo-Soo Lee, Yang-Bok Lee, Soo-Mi Choi, Yong-Guk Kim,
Moon-Hyun Kim*

Probabilistic Face Tracking Using Boosted Multi-view Detector	577
<i>Peihua Li, Haijing Wang</i>	
Face Samples Re-lighting for Detection Based on the Harmonic Images	585
<i>Jie Chen, Yuemin Li, Laiyun Qing, Baocai Yin, Wen Gao</i>	
A Dictionary Registration Method for Reducing Lighting Fluctuations in Subspace Face Recognition	593
<i>Kenji Matsuo, Masayuki Hashimoto, Atsushi Koike</i>	
Applications (I)	
A 3D-Dialogue System Between Game Characters to Improve Reality in MMORPG	601
<i>Dae-Woong Rhee, Il-Seok Won, Hyunjoo Song, Hung Kook Park, Juno Chang, Kang Ryoung Park, Yongjoo Cho</i>	
A Hybrid Approach to Detect Adult Web Images	609
<i>Qing-Fang Zheng, Ming-Ji Zhang, Wei-Qiang Wang</i>	
A Hierarchical Dynamic Bayesian Network Approach to Visual Tracking	617
<i>Hua Li, Rong Xiao, Hong-Jiang Zhang, Li-Zhong Peng</i>	
A Hybrid Architectural Framework for Digital Home Multimedia Multi-modal Collaboration Services	625
<i>Doo-Hyun Kim, Vinod Cherian Joseph, Kyunghee Lee, Eun Hwan Jo</i>	
E-learning as Computer Games: Designing Immersive and Experiential Learning	633
<i>Ang Chee Siang, G.S.V. Radha Krishna Rao</i>	
Event-Based Surveillance System for Efficient Monitoring	641
<i>Do Joon Jung, Se Hyun Park, Hang Joon Kim</i>	
A Collaborative Multimedia Authoring System Based on the Conceptual Temporal Relations	649
<i>Mee Young Sung</i>	
Multimedia Integration for Cooking Video Indexing	657
<i>Reiko Hamada, Koichi Miura, Ichiro Ide, Shin'ichi Satoh, Shuichi Sakai, Hidehiko Tanaka</i>	
Teleconference System with a Shared Working Space and Face Mouse Interaction	665
<i>Jin Hak Kim, Sang Chul Ahn, Hyoung-Gon Kim</i>	

User Interface (II)

Performance Analysis for Serially Concatenated FEC in IEEE802.16a over Wireless Channels	672
<i>Kao-Lung Huang, Hsueh-Ming Hang</i>	
Successive Interference Cancellation for CDMA Wireless Multimedia Services	680
<i>Jin Young Kim, Yong Kim</i>	
SCORM-Based Contents Collecting Using Mobile Agent in M-Learning	688
<i>Sun-Gwan Han, Hee-Seop Han, Jae-Bong Kim</i>	
Improved Bit-by-Bit Binary Tree Algorithm in Ubiquitous ID System	696
<i>Ho-Seung Choi, Jae-Ryong Cha, Jae-Hyun Kim</i>	
Automatic Synchronized Browsing of Images Across Multiple Devices	704
<i>Zhigang Hua, Xing Xie, Hanqing Lu, Wei-Ying Ma</i>	
An Intelligent Handoff Protocol for Adaptive Multimedia Streaming Service in Mobile Computing Environment	712
<i>Jang-Woon Baek, Dae-Wha Seo</i>	
Platform Architecture for Seamless MMS Service over WLAN and CDMA2000 Networks	720
<i>Su-Yong Kim, Yong-Bum Cho, Sung-Joon Cho</i>	
Make Stable QoS in Wireless Multimedia Ad Hoc Network with Transmission Diversity	728
<i>Chao Zhang, Mingmei Li, Xiaokang Lin, Shigeki Yamada, Mitsutoshi Hatori</i>	
Gaze from Motion: Towards Natural User Interfaces	736
<i>Mun-Ho Jeong, Masamichi Ohsugi, Ryuji Funayama, Hiroki Mori</i>	
The Development of MPEG-4 Based RTSP System for Mobile Multimedia Streaming Services	746
<i>Sangeun Lee, Hyunwoo Park, Taesoo Yun</i>	
Seamless Mobile Service for Pervasive Multimedia	754
<i>Enyi Chen, Degan Zhang, Yuanchun Shi, Guangyou Xu</i>	
Transformation of MPEG-4 Contents for a PDA Device	762
<i>Sangwook Kim, Kyungdeok Kim, Sookyoung Lee</i>	

Semantic Retrieval in a Large-Scale Video Database by Using Both Image and Text Feature	770
<i>Chuan Yu, Hiroshi Mo, Norio Katayama, Shin'ichi Satoh, Shoichiro Asano</i>	
 Image Analysis (II)	
Performance of Correlation-Based Stereo Algorithm with Respect to the Change of the Window Size	778
<i>Dong-Min Woo, Howard Schultz, Edward Riseman, Allen Hanson</i>	
Consideration of Illuminant Independence in MPEG-7 Color Descriptors	786
<i>Sang-Kyun Kim, Yanglim Choi, Wonhee Choe, Du-Sik Park, Ji-Yeun Kim, Yang-Seock Seo</i>	
Improvement on Colorization Accuracy by Partitioning Algorithm in CIELAB Color Space	794
<i>Tomohisa Takahama, Takahiko Horiuchi, Hiroaki Kotera</i>	
Gabor-Kernel Fisher Analysis for Face Recognition	802
<i>Baochang Zhang</i>	
Film Line Scratch Detection Using Neural Network	810
<i>Sin Kuk Kang, Eun Yi Kim, Kee Chul Jung, Hang Joon Kim</i>	
A Simplified Half Pixel Motion Estimation Algorithm Based on the Spatial Correlation	818
<i>HyoSun Yoon, Miyoung Kim</i>	
A New Tracking Mechanism for Semi-automatic Video Object Segmentation	824
<i>Zhi Liu, Jie Yang, Ningsong Peng</i>	
A Visual Model for Estimating the Perceptual Redundancy Inherent in Color Images	833
<i>Chun-Hsien Chou, Kuo-Cheng Liu</i>	
Selective Image Sharpening by Simultaneous Nonlinear-Diffusion Process with Spatially Varying Parameter Presetting	841
<i>Takahiro Saito, Shigemitsu Anyoji, Takashi Komatsu</i>	
Directional Weighting-Based Demosaicking Algorithm	849
<i>Tsung-Nan Lin, Chih-Lung Hsu</i>	
A New Text Detection Algorithm in Images/Video Frames	858
<i>Qixiang Ye, Qingming Huang</i>	

Automatic Video Object Tracking Using a Mosaic-Based Background 866
Young-Kee Jung, Kyu-Won Lee, Dong-Min Woo, Yo-Sung Ho

Audio Analysis

Semantic Region Detection in Acoustic Music Signals 874
Namunu Chinthaka Maddage, Changsheng Xu, Arun Shenoy, Ye Wang

Audio Classification for Radio Broadcast Indexing:
 Feature Normalization and Multiple Classifiers Decision 882
Christine S enac, Eliathamby Ambikairajh

Dominant Feature Vectors Based Audio Similarity Measure 890
Jing Gu, Lie Lu, Rui Cai, Hong-Jiang Zhang, Jian Yang

**Moving from Content
 to Concept-Based Image/Video Retrieval**

Generative Grammar of Elemental Concepts 898
Jose A. Lay, Ling Guan

Learning Image Manifold Using Web Data 907
Xin-Jing Wang, Wei-Ying Ma, Xing Li

Novel Concept for Video Retrieval in Life Log Application 915
Datchakorn Tancharoen, Kiyoharu Aizawa

Content-Based Retrieval of 3D Head Models
 Using a Single Face View Query 924
Pui Fong Yeung, Hau San Wong, Horace H.-S. Ip

Region-Based Image Retrieval with Perceptual Colors 931
Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma

H.264

Converting DCT Coefficients to H.264/AVC Transform Coefficients 939
Jun Xin, Anthony Vetro, Huifang Sun

An Adaptive Hybrid Mode Decision Scheme for H.264/AVC Video
 over Unreliable Packet Networks 947
Feng Huang, Jenq-Neng Hwang, Yuzhuo Zhong

Workload Characterization of the H.264/AVC Decoder 957
Tse-Tsung Shih, Chia-Lin Yang, Yi-Shin Tung

An Efficient Traffic Smoothing Method for MPEG-4 Part-10 AVC/H.264 Bitstream over Wireless Network	967
<i>Kwang-Pyo Choi, Inchoon Choi, Byeungwoo Jeon, Keun-Young Lee</i>	
An Error Resilience Scheme for Packet Loss Recover of H.264 Video	975
<i>Ziqing Mao, Rong Yan, Ling Shao, Dong Xie</i>	
Key Techniques of Bit Rate Reduction for H.264 Streams	985
<i>Peng Zhang, Qing-Ming Huang, Wen Gao</i>	
Face, Gesture, and Behavior (III)	
Salient Region Detection Using Weighted Feature Maps Based on the Human Visual Attention Model	993
<i>Yiqun Hu, Xing Xie, Wei-Ying Ma, Liang-Tien Chia, Deepu Rajan</i>	
An Attention-Based Decision Fusion Scheme for Multimedia Information Retrieval	1001
<i>Xian-Sheng Hua, Hong-Jiang Zhang</i>	
Approximating Inference on Complex Motion Models Using Multi-model Particle Filter	1011
<i>Jianyu Wang, Debin Zhao, Shiguang Shan, Wen Gao</i>	
Human Activity Recognition in Archaeological Sites by Hidden Markov Models	1019
<i>Marco Leo, Paolo Spagnolo, Tiziana D’Orazio, Arcangelo Distanto</i>	
An HMM Based Gesture Recognition for Perceptual User Interface	1027
<i>HySun Park, EunYi Kim, SangSu Jang, HangJoon Kim</i>	
Vision-Based Sign Language Recognition Using Sign-Wise Tied Mixture HMM	1035
<i>Liangguo Zhang, Gaolin Fang, Wen Gao, Xilin Chen, Yiqiang Chen</i>	
Author Index	1043

Table of Contents, Part I

Volume I

Art

Categorizing Traditional Chinese Painting Images	1
<i>Shuqiang Jiang, Tiejun Huang</i>	
A Knowledge-Driven Approach for Korean Traditional Costume (Hanbok) Modeling	9
<i>Yang-Hee Nam, Bo-Ran Lee, Crystal S. Oh</i>	
Retrieval of Chinese Calligraphic Character Image	17
<i>Yueting Zhuang, Xiafen Zhang, Jiangqin Wu, Xiqun Lu</i>	

Network (I)

Random Channel Allocation Scheme in HIPERLAN/2	25
<i>Eui-Seok Hwang, Jeong-Jae Won, You-Chang Ko, Hyong-Woo Lee, Choong-Ho Cho</i>	
A Two-Stage Queuing Approach to Support Real-Time QoS Guarantee for Multimedia Services in TDMA Wireless Networks	33
<i>Ren-Hao Cheng, Po-Cheng Huang, Mong-Fong Horng, Jiang-Shiung Ker, Yau-Hwang Kuo</i>	
Performance Evaluation of Adaptive Rate Control (ARC) for Multimedia Traffic	41
<i>Surasee Prahmkaew, Chanintorn Jittawiriyankoon</i>	

Sports (I)

A Tennis Video Indexing Approach Through Pattern Discovery in Interactive Process	49
<i>Peng Wang, Rui Cai, Shi-Qiang Yang</i>	
Online Play Segmentation for Broadcasted American Football TV Programs	57
<i>Liexian Gu, Xiaoqing Ding, Xian-Sheng Hua</i>	
Semantic Analysis of Basketball Video Using Motion Information	65
<i>Song Liu, Haoran Yi, Liang-Tien Chia, Deepu Rajan, Syin Chan</i>	

Immersive Conferencing: Novel Interfaces and Paradigms for Remote Collaboration

Reach-Through-the-Screen: A New Metaphor for Remote Collaboration	73
<i>Jonathan Foote, Qiong Liu, Don Kimber, Patrick Chiu, Frank Zhao</i>	
Working Documents	81
<i>Paul Luff, Hideaki Kuzuoka, Christian Heath, Jun Yamashita, Keiichi Yamazaki</i>	
Immersive Meeting Point	89
<i>Ralf Tanger, Peter Kauff, Oliver Schreer</i>	
EnhancedTable: Supporting a Small Meeting in Ubiquitous and Augmented Environment	97
<i>Hideki Koike, Shin'ichiro Nagashima, Yasuto Nakanishi, Yoichi Sato</i>	
Remote Collaboration on Physical Whiteboards	105
<i>Zhengyou Zhang, Li-wei He</i>	

Network (II)

Aggressive Traffic Smoothing for Delivery of Online Multimedia	114
<i>Jeng-Wei Lin, Ray-I Chang, Jan-Ming Ho, Feipei Lai</i>	
Distributed Video Streaming Using Multicast (DVSM)	122
<i>Ramesh Yerraballi, ByungHo Lee</i>	
Centralized Peer-to-Peer Streaming with PFGS Video Codec	131
<i>Ivan Lee, Ling Guan</i>	
Buffer Level Estimation for Seamless Media Streaming in Mobile IPv6 Networks	139
<i>Dongwook Lee, JongWon Kim</i>	
Dynamic Walks for Searching Streaming Media in Peer-to-Peer Networks	147
<i>Zhou Su, Jiro Katto, Yasuhiko Yasuda</i>	

Image Retrieval

An Image Retrieval Scheme Using Multi-instance and Pseudo Image Concepts	157
<i>Feng-Cheng Chang, Hsueh-Ming Hang</i>	

Representation of Clipart Image Using Shape and Color with Spatial Relationship	165
<i>Jeong-Hyun Cho, Chang-Gyu Choi, Yongseok Chang, Sung-Ho Kim</i>	
Automatic Categorization for WWW Images with Applications for Retrieval Navigation	174
<i>Koji Nakahira, Satoshi Ueno, Kiyoharu Aizawa</i>	
Region-Based Image Retrieval with Scale and Orientation Invariant Features	182
<i>Surong Wang, Liang-Tien Chia, Deepu Rajan</i>	
SOM-Based Sample Learning Algorithm for Relevance Feedback in CBIR	190
<i>Tatsunori Nishikawa, Takahiko Horiuchi, Hiroaki Kotera</i>	
Classification of Digital Photos Taken by Photographers or Home Users	198
<i>Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, Changshui Zhang</i>	
Image Analysis (I)	
Background Modeling Using Phase Space for Day and Night Video Surveillance Systems	206
<i>Yu-Ming Liang, Arthur Chun-Chieh Shih, Hsiao-Rong Tyan, Hong-Yuan Mark Liao</i>	
Sequential Robust Direct Motion Estimation with Equal Projective Basis	214
<i>Jong-Eun Ha, Dong-Joong Kang, Muh-Ho Jeong</i>	
Generation of 3D Urban Model Using Cooperative Hybrid Stereo Matching.....	222
<i>Dong-Min Woo, Howard Schultz, Young-Kee Jung, Kyu-Won Lee</i>	
Segmentation of Interest Objects Using the Hierarchical Mesh Structure	230
<i>Dong-Keun Lim, Yo-Sung Ho</i>	
A Rapid Scheme for Slow-Motion Replay Segment Detection	239
<i>Wei-Hong Chuang, Dun-Yu Hsiao, Soo-Chang Pei, Homer Chen</i>	
Recognition of Very Low-Resolution Characters from Motion Images Captured by a Portable Digital Camera	247
<i>Shinsuke Yanadume, Yoshito Mekada, Ichiro Ide, Hiroshi Murase</i>	

An Effective Anchorperson Shot Extraction Method Robust to False Alarms	255
<i>Sang-Kyun Kim, Doo Sun Hwang, Ji-Yeun Kim, Yang-Seock Seo</i>	
A Region Based Image Matching Method with Regularized SAR Model	263
<i>Yaowei Wang, Weiqiang Wang, Yanfei Wang</i>	
Shot Classification and Scene Segmentation Based on MPEG Compressed Movie Analysis	271
<i>Masaru Sugano, Masanori Furuya, Yasuyuki Nakajima, Hiromasa Yanagihara</i>	
Moving Object Segmentation: A Block-Based Moving Region Detection Approach	280
<i>Wei Zeng, Qingming Huang</i>	
A Multi-view Camera Tracking for Modeling of Indoor Environment	288
<i>Kiyoung Kim, Woontack Woo</i>	
Image Registration Using Triangular Mesh	298
<i>Ben Yip, Jesse S. Jin</i>	
Online Learning Objectionable Image Filter Based on SVM	304
<i>Yang Liu, Wei Zeng, Hongxun Yao</i>	
Motion Objects Segmentation Using a New Level Set Based Method	312
<i>Hongqiang Bao, Zhaoyang Zhang</i>	
3-D Shape Analysis of Anatomical Structures Based on an Interactive Multiresolution Approach	319
<i>Soo-Mi Choi, Jeong-Sik Kim, Yong-Guk Kim, Joo-Young Park</i>	
Integrating Color, Texture, and Spatial Features for Image Interpretation	327
<i>Hui-Yu Huang, Yung-Sheng Chen, Wen-Hsing Hsu</i>	
Automatic Video Genre Detection for Content-Based Authoring	335
<i>Sung Ho Jin, Tae Meon Bae, Yong Man Ro</i>	
Face, Gesture, and Behavior (I)	
Face Appeal Model Based on Statistics	344
<i>Bi Song, Mingjing Li, Zhiwei Li, Hong-Jiang Zhang, Zhengkai Liu</i>	
A Novel Gabor-LDA Based Face Recognition Method	352
<i>Yanwei Pang, Lei Zhang, Mingjing Li, Zhengkai Liu, Weiyang Ma</i>	

Gesture-Based User Interfaces for Handheld Devices Using Accelerometer	359
<i>Ikjin Jang, Wonbae Park</i>	
Face and Gesture Recognition Using Subspace Method for Human-Robot Interaction	369
<i>Md. Hasanuzzaman, T. Zhang, V. Ampornaramveth, M.A. Bhuiyan, Y. Shirai, H. Ueno</i>	
Spatial Histogram Features for Face Detection in Color Images	377
<i>Hongming Zhang, Debin Zhao</i>	
Correlation Filters for Facial Recognition Login Access Control	385
<i>Daniel E. Riedel, Wanquan Liu, Ronny Tjahyadi</i>	
Easy and Convincing Ear Modeling for Virtual Human	394
<i>Hui Zhang, In Kyu Park</i>	
Virtual Reality and Computer Graphics	
A Polyhedral Object Recognition Algorithm for Augmented Reality	402
<i>Dong-Joong Kang, Jong-Eun Ha, Muh-Ho Jeong</i>	
Spectral Coding of Three-Dimensional Mesh Geometry Information Using Dual Graph	410
<i>Sung-Yeol Kim, Seung-Uk Yoon, Yo-Sung Ho</i>	
Real-Time Free-Viewpoint Video Generation Using Multiple Cameras and a PC-Cluster	418
<i>Megumu Ueda, Daisaku Arita, Rin-ichiro Taniguchi</i>	
Framework for Smooth Optical Interaction Using Adaptive Subdivision in Virtual Production	426
<i>Seung Man Kim, Naveen Dachuri, Kwan H. Lee</i>	
Projection-Based Registration Using Color and Texture Information for Virtual Environment Generation	434
<i>Sehwan Kim, Kiyoun Kim, Woontack Woo</i>	
Enhanced Synergistic Image Creator: An NPR Method with Natural Curly Brushstrokes	444
<i>Atsushi Kasao, Kazunori Miyata</i>	
Content Production (I)	
An XMT Authoring System Supporting Various Presentation Environments	453
<i>Heesun Kim</i>	

An Adaptive Scene Compositor Model in MPEG-4 Player for Mobile Device	461
<i>Hyunju Lee, Sangwook Kim</i>	
JPEG2000 Image Adaptation for MPEG-21 Digital Items	470
<i>Yiqun Hu, Liang-Tien Chia, Deepu Rajan</i>	
An Annotation Method and Application for Video Contents Based on a Semantic Graph	478
<i>Tomohisa Akafuji, Kaname Harumoto, Keishi Kandori, Kôiti Hasida, Shinji Shimojo</i>	
An Annotated-Objects Assist Method for Extraction of Ordinary-Objects in a Video Content Generation Support System	487
<i>Wenli Zhang, Shunsuke Kamijyo, Masao Sakauchi</i>	
Intelligent Media Integration for Social Information Infrastructure	
Free-Viewpoint TV (FTV) System	497
<i>Toshiaki Fujii, Masayuki Tanimoto</i>	
In-Car Speech Recognition Using Distributed Multiple Microphones	505
<i>Weifeng Li, Takanori Nishino, Chiyomi Miyajima, Katsunobu Itou, Kazuya Takeda, Fumitada Itakura</i>	
Development of Advanced Image Processing Technology and Its Application to Computer Assisted Diagnosis and Surgery	514
<i>Takayuki Kitasaka, Kensaku Mori, Yasuhito Suenaga</i>	
Discussion Mining: Annotation-Based Knowledge Discovery from Real World Activities	522
<i>Katashi Nagao, Katsuhiko Kaji, Daisuke Yamamoto, Hironori Tomobe</i>	
Determining Correspondences Between Sensory and Motor Signals	532
<i>Kento Nishibori, Jinji Chen, Yoshinori Takeuchi, Tetsuya Matsumoto, Hiroaki Kudo, Noboru Ohnishi</i>	
Approaches or Methods of Security Engineering	
Architecture of Authentication Mechanism for Emerging T-commerce Environments	540
<i>Sangkyun Kim, Hong Joo Lee, Choon Seong Leem</i>	
Threat Description for Developing Security Countermeasure	548
<i>Seung-youn Lee, Myong-chul Shin, Jae-sang Cha, Tai-hoon Kim</i>	

RPS: An Extension of Reference Monitor to Prevent Race-Attacks	556
<i>Jongwoon Park, Gunhee Lee, Sangha Lee, Dong-kyoo Kim</i>	
SITIS: Scalable Intrusion Tolerance Middleware for Internet Service Survivability	564
<i>GangShin Lee, Chaetae Im, TaeJin Lee, HyungJong Kim, Dong Hoon Lee</i>	
On Minimizing Distortion in Secure Data-Hiding for Binary Images	572
<i>Yongsu Park, Yookun Cho</i>	
Software Design Method Enhanced by Appended Security Requirements	578
<i>Eun-ser Lee, Sun-myoung Hwang</i>	

Multimedia Servers

On the Disk Layout for Periodical Video Broadcast Services	586
<i>Meng-Huang Lee</i>	
Paged Segment Striping Scheme for the Dynamic Scheduling of Continuous Media Servers	594
<i>KyungOh Lee, J.B. Lee, Kee-Wook Rim</i>	
A Heuristic Packet Scheduling Algorithm for Media Streaming with Server Diversity	602
<i>Yajie Liu, Wenhua Dou, Heying Zhang</i>	
Proxy Caching Scheme Based on the Characteristics of Streaming Media Contents on the Internet	610
<i>Hyeonok Hong, Seungwon Lee, Seongho Park, Yongju Kim, Kidong Chung</i>	
A Distributed VOD Server Based on VIA and Interval Cache	618
<i>Soo-Cheol Oh, Sang-Hwa Chung</i>	

Video Retrieval

Clustering of Video Packets Using Interactive Refinement by Relevance Feedback	626
<i>Yukihiro Kinoshita, Naoko Nitta, Noboru Babaguchi</i>	
Semantic Video Indexing and Summarization Using Subtitles	634
<i>Haoran Yi, Deepu Rajan, Liang-Tien Chia</i>	
Audio Visual Cues for Video Indexing and Retrieval	642
<i>Paisarn Muneesawang, Tahir Amin, Ling Guan</i>	

Key Image Extraction from a News Video Archive for Visualizing Its Semantic Structure	650
<i>Hiroshi Mo, Fuminori Yamagishi, Ichiro Ide, Norio Katayama, Shin'ichi Satoh, Masao Sakauchi</i>	
Author Index	659

Table of Contents, Part III

Human-Scale Virtual Reality and Interaction

WARAJI: Foot-Driven Navigation Interfaces for Virtual Reality Applications	1
<i>Salvador Barrera, Piperakis Romanos, Suguru Saito, Hiroki Takahashi, Masayuki Nakajima</i>	
Time Space Interface Using DV (Digital Video) and GPS (Global Positioning System) Technology – A Study with an Art Project “Field-Work@Alsace”	8
<i>Masaki Fujihata</i>	
Action Generation from Natural Language	15
<i>Satoshi Funatsu, Tomofumi Koyama, Suguru Saito, Takenobu Tokunaga, Masayuki Nakajima</i>	
Human-Scale Interaction with a Multi-projector Display and Multimodal Interfaces	23
<i>Naoki Hashimoto, Jaeho Ryu, Seungzoo Jeong, Makoto Sato</i>	
Entertainment Applications of Human-Scale Virtual Reality Systems	31
<i>Akihiko Shirai, Kiichi Kobayashi, Masahiro Kawakita, Shoichi Hasegawa, Masayuki Nakajima, Makoto Sato</i>	
Analysis and Synthesis of Latin Dance Using Motion Capture Data	39
<i>Noriko Nagata, Kazutaka Okumoto, Daisuke Iwai, Felipe Toro, Seiji Inokuchi</i>	

Surveillance and Tracking

Web-Based Telepresence System Using Omni-directional Video Streams	45
<i>Kazumasa Yamazawa, Tomoya Ishikawa, Tomokazu Sato, Sei Ikeda, Yutaka Nakamura, Kazutoshi Fujikawa, Hideki Sunahara, Naokazu Yokoya</i>	
Wide View Surveillance System with Multiple Smart Image Sensors and Mirrors	53
<i>Ryusuke Kawahara, Satoshi Shimizu, Takayuki Hamamoto</i>	
Object Tracking and Identification in Video Streams with Snakes and Points	61
<i>Bruno Lameyre, Valerie Gouet</i>	

Optical Flow-Based Tracking of Deformable Objects Using a Non-prior Training Active Feature Model	69
<i>Sangjin Kim, Jinyoung Kang, Jeongho Shin, Seongwon Lee, Joonki Paik, Sangkyu Kang, Besma Abidi, Mongi Abidi</i>	
An Immunological Approach to Raising Alarms in Video Surveillance	79
<i>Lukman Sasmita, Wanquan Liu, Svetha Venkatesh</i>	
Sat-Cam: Personal Satellite Virtual Camera	87
<i>Hansung Kim, Itaru Kitahara, Kiyoshi Kogure, Norihiro Hagita, Kwanghoon Sohn</i>	

Image Analysis (III)

A Linear Approximation Based Method for Noise-Robust and Illumination-Invariant Image Change Detection	95
<i>Bin Gao, Tie-Yan Liu, Qian-Sheng Cheng, Wei-Ying Ma</i>	
3D Model Similarity Measurement with Geometric Feature Map Based on Phase-Encoded Range Image	103
<i>Donghui Wang, Chenyang Cui</i>	
Automatic Peak Number Detection in Image Symmetry Analysis	111
<i>Jingrui He, Mingjing Li, Hong-Jiang Zhang, Hanghang Tong, Changshui Zhang</i>	
Image Matching Based on Singular Value Decomposition	119
<i>Feng Zhao</i>	
Image Matching Based on Scale Invariant Regions	127
<i>Lei Qin, Wei Zeng, Weiqiang Wang</i>	

Compression (II)

A Method for Blocking Effect Reduction Based on Optimal Filtering	135
<i>Daehee Kim, Yo-Sung Ho</i>	
Novel Video Error Concealment Using Shot Boundary Detection	143
<i>You-Neng Xiao, Xiang-Yang Xue, Ruo-Nan Pu, Hong Lu, Congjie Mi</i>	
Using Only Long Windows in MPEG-2/4 AAC Encoding	151
<i>Fu-Mau Chang, Shingchern D. You</i>	
Frequency Weighting and Selective Enhancement for MPEG-4 Scalable Video Coding	159
<i>Seung-Hwan Kim, Yo-Sung Ho</i>	
Efficient Multiview Video Coding Based on MPEG-4	167
<i>Wenxian Yang, King Ngi Ngan, Jianfei Cai</i>	

Block Matching Using Integral Frame Attributes	175
<i>Viet Anh Nguyen, Yap-Peng Tan</i>	
Impact of Similarity Threshold on Arbitrary Shaped Pattern Selection Very Low Bit-Rate Video Coding Algorithm	184
<i>Manoranjan Paul, Manzur Murshed</i>	
A Study on the Quantization Scheme in H.264/AVC and Its Application to Rate Control	192
<i>Siwei Ma, Wen Gao, Debin Zhao, Yan Lu</i>	
An Efficient VLSI Implementation for MC Interpolation of AVS Standard	200
<i>Lei Deng, Wen Gao, Ming-Zeng Hu, Zhen-Zhou Ji</i>	
Fast Fractal Image Encoder Using Non-overlapped Block Classification and Simplified Isometry Testing Scheme	207
<i>Youngjoon Han, Hawik Chung, Hernsoo Hahn</i>	
A Fast Downsizing Video Transcoder Based on H.264/AVC Standard	215
<i>Chih-Hung Li, Chung-Neng Wang, Tihao Chiang</i>	
Temporal Error Concealment with Block Boundary Smoothing	224
<i>Woong Il Choi, Byeungwoo Jeon</i>	
Spatio-temporally Adaptive Regularization for Enhancement of Motion Compensated Wavelet Coded Video	232
<i>Junghoon Jung, Hyunjong Ki, Seongwon Lee, Jeongho Shin, Jinyoung Kang, Joonki Paik</i>	
ROI and FOI Algorithms for Wavelet-Based Video Compression	241
<i>Chaoqiang Liu, Tao Xia, Hui Li</i>	
Adaptive Distributed Source Coding for Multi-view Images	249
<i>Mehrdad Panahpour Tehrani, Michael Droese, Toshiaki Fujii, Masayuki Tanimoto</i>	
Hybrid Multiple Description Video Coding Using SD/MD Switching	257
<i>Il Koo Kim, Nam Ik Cho</i>	
Streaming (II)	
Clusters-Based Distributed Streaming Services with Fault-Tolerant Schemes	265
<i>Xiaofei Liao, Hai Jin</i>	
Towards SMIL Document Analysis Using an Algebraic Time Net	273
<i>A. Abdelli, M. Daoudi</i>	

MULTFRC-LERD: An Improved Rate Control Scheme for Video Streaming over Wireless	282
<i>Xiaolin Tong, Qingming Huang</i>	
Multi-source Media Streaming for the Contents Distribution in a P2P Network	290
<i>Sung Yong Lee, Jae Gil Lee, Chang Yeol Choi</i>	
A New Feature for TV Programs: Viewer Participation Through Videoconferencing	298
<i>Jukka Rauhala, Petri Vuorimaa</i>	
Application Layer Multicast with Proactive Route Maintenance over Redundant Overlay Trees	306
<i>Yohei Kunichika, Jiro Katto, Sakae Okubo</i>	
Real-Time Rate Control Via Variable Frame Rate and Quantization Parameters	314
<i>Chi-Wah Wong, Oscar C. Au, Raymond Chi-Wing Wong, Hong-Kwai Lam</i>	
The Structure of Logically Hierarchical Cluster for the Distributed Multimedia on Demand	323
<i>Xuhui Xiong, Shengsheng Yu, Jingli Zhou</i>	
Watermarking (II)	
Image Forensics Technology for Digital Camera	331
<i>Jongweon Kim, Youngbae Byun, Jonguk Choi</i>	
Lossless Data Hiding Based on Histogram Modification of Difference Images	340
<i>Sang-Kwang Lee, Young-Ho Suh, Yo-Sung Ho</i>	
A Robust Image Watermarking Scheme to Geometrical Attacks for Embedment of Multibit Information	348
<i>Jung-Soo Lee, Whoi-Yul Kim</i>	
Combined Encryption and Watermarking Approaches for Scalable Multimedia Coding	356
<i>Feng-Cheng Chang, Hsiang-Cheh Huang, Hsueh-Ming Hang</i>	
Digital Image Watermarking Using Independent Component Analysis . . .	364
<i>Viet Thang Nguyen, Jagdish Chandra Patra</i>	
Clustering-Based Image Retrieval Using Fast Exhaustive Multi-resolution Search Algorithm	372
<i>Byung Cheol Song, Kang Wook Chun</i>	

Robust Watermarking for Copyright Protection of 3D Polygonal Model	378
<i>Wan-Hyun Cho, Myung-Eun Lee, Hyun Lim, Soon-Young Park</i>	
Digital Video Scrambling Method Using Intra Prediction Mode	386
<i>Jinhaeng Ahn, Hiuk Jae Shim, Byeungwoo Jeon, Inchoon Choi</i>	
Watermark Re-synchronization Using Sinusoidal Signals in DT-CWT Domain	394
<i>Miin-Luen Day, Suh-Yin Lee, I-Chang Jou</i>	
The Undeniable Multi-signature Scheme Suitable for Joint Copyright Protection on Digital Contents	402
<i>Sung-Hyun Yun, Hyung-Woo Lee</i>	
A Digital Watermarking Scheme for Personal Image Authentication Using Eigenface	410
<i>Chien-Hsung Chen, Long-Wen Chang</i>	
Cryptanalysis of a Chaotic Neural Network Based Multimedia Encryption Scheme	418
<i>Chengqing Li, Shujun Li, Dan Zhang, Guanrong Chen</i>	
A Secure Steganographic Method on Wavelet Domain of Palette-Based Images	426
<i>Wei Ding, Xiang-Wei Kong, Xin-Gang You, Zi-Ren Wang</i>	
Mapping Energy Video Watermarking Algorithm Based on Compressed Domain	433
<i>Lijun Wang, HongXun Yao, ShaoHui Liu, Wen Gao</i>	
A Fragile Image Watermarking Based on Mass and Centroid	441
<i>Hui Han, HongXun Yao, ShaoHui Liu, Yan Liu</i>	
Content Production (II)	
MPEG-21 DIA Testbed for Stereoscopic Adaptation of Digital Items	449
<i>Hyunsik Sohn, Haksoo Kim, Manbae Kim</i>	
An MPEG-4 Authoring System with Temporal Constraints for Interactive Scene	457
<i>Heesun Kim</i>	
A Method of Digital Camera Work Focused on Players and a Ball (– Toward Automatic Contents Production System of Commentary Soccer Video by Digital Shooting –)	466
<i>Masahito Kumano, Yasuo Arika, Kiyoshi Tsukada</i>	
Real-Time Rendering of Watercolor Effects for Virtual Environments	474
<i>Su Ian Eugene Lei, Chun-Fa Chang</i>	

Haptic Interaction in Realistic Multimedia Broadcasting	482
<i>Jongeun Cha, Jeha Ryu, Seungjun Kim, Seongeun Eom, Byungha Ahn</i>	

Object-Based Stereoscopic Conversion of MPEG-4 Encoded Data	491
<i>Manbae Kim, Sanghoon Park, Youngran Cho</i>	

Applications (II)

Shared Annotation Database for Networked Wearable Augmented Reality System	499
<i>Koji Makita, Masayuki Kanbara, Naokazu Yokoya</i>	

A Study on Image Electronic Money Based on Watermarking Technique	508
<i>Jung-Soo Lee, Jong-Weon Kim, Kyu-Tae Kim, Jong-Uk Choi, Whoi-Yul Kim</i>	

A Fully Automated Web-Based TV-News System	515
<i>P.S. Lai, L.Y. Lai, T.C. Tseng, Y.H. Chen, Hsin-Chia Fu</i>	

An Evolutionary Computing Approach for Mining of Bio-medical Images	523
<i>Shashikala Tapaswi, R.C. Joshi</i>	

Movie-Based Multimedia Environment for Programming and Algorithms Design	533
<i>Dmitry Vazhenin, Alexander Vazhenin, Nikolay Mirenkov</i>	

An Improvement Algorithm for Accessing Patterns Through Clustering in Interactive VRML Environments	542
<i>Damon Shing-Min Liu, Shao-Shin Hung, Ting-Chia Kuo</i>	

Multimedia Analysis

MPEG-4 Video Retrieval Using Video-Objects and Edge Potential Functions	550
<i>Minh-Son Dao, Francesco G.B. DeNatale, Andrea Massa</i>	

A Unified Framework Using Spatial Color Descriptor and Motion-Based Post Refinement for Shot Boundary Detection	558
<i>Wei-Ta Chu, Wen-Huang Cheng, Sheng-Fang He, Chia-Wei Wang, Ja-Ling Wu</i>	

HMM-Based Audio Keyword Generation	566
<i>Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, Qi Tian</i>	

Video Scene Segmentation Using Sequential Change Detection	575
<i>Zhenyan Li, Hong Lu, Yap-Peng Tan</i>	
Feature Extraction and Evaluation Using Edge Histogram Descriptor in MPEG-7	583
<i>Chee Sun Won</i>	
Automatic Synthesis of Background Music Track Data by Analysis of Video Contents	591
<i>Toshio Modegi</i>	
Compression (III)	
Picture Quality Improvement in Low Bit Rate Video Coding Using Block Boundary Classification and Simple Adaptive Filter	599
<i>Keek-Koo Kwon, Jin-Suk Ma, Sung-Ho Im, Dong-Sun Lim</i>	
Bit Position Quantization in Scalable Video Coding for Representing Detail of Image	607
<i>Hideaki Kimata, Masaki Kitahara, Kazuto Kamikura, Yoshiyuki Yashima</i>	
A Fast Full Search Algorithm for Multiple Reference Images	615
<i>Hyun-Soo Kang, Si-Woong Lee, Kook-Yeol Yoo, Jae-Gark Choi</i>	
Preprocessing of Depth and Color Information for Layered Depth Image Coding	622
<i>Seung-Uk Yoon, Sung-Yeol Kim, Yo-Sung Ho</i>	
Selective Motion Estimation for Fast Video Encoding	630
<i>Sun Young Lee, Yong Ho Cho, Whoiyul Kim, Euee S. Jang</i>	
An Efficient VLSI Architecture of the Sample Interpolation for MPEG-4 Advanced Simple Profile	639
<i>Lei Deng, Ming-Zeng Hu, Zhen-Zhou Ji</i>	
Performance Improvement of Vector Quantization by Using Threshold . . .	647
<i>Hung-Yi Chang, Pi-Chung Wang, Rong-Chang Chen, Shuo-Cheng Hu</i>	
An Efficient Object Based Personal Video Coding System	655
<i>Cataldo Guaragnella, Tiziana D' Orazio</i>	
Multiview Video Coding Based on Global Motion Model	665
<i>Xun Guo, Qingming Huang</i>	
A Novel Rate-Distortion Model for Leaky Prediction Based FGS Video Coding	673
<i>Jianhua Wu, Jianfei Cai</i>	

JPEG Quantization Table Design for Photos with Face in Wireless Handset	681
<i>Gu-Min Jeong, Jun-Ho Kang, Yong-Su Mun, Doo-Hee Jung</i>	
Effective Drift Reduction Technique for Reduced Bit-Rate Video Adaptation	689
<i>June-Sok Lee, Goo-Rak Kwon, Jae-Won Kim, Jae-Yong Lee, Sung-Jea Ko</i>	
On Implementation of a Scalable Wallet-Size Cluster Computing System for Multimedia Applications	697
<i>Liang-Teh Lee, Kuan-Ching Li, Chao-Tung Yang, Chia-Ying Tseng, Kang-Yuan Liu, Chih-Hung Hung</i>	
An Adaptive LS-Based Motion Prediction Algorithm for Video Coding	705
<i>Min-Cheol Hong, Myung-Sik Yoo, Ji-Hee Kim</i>	
Embedded Packetization Framework for Layered Multiple Description Coding	713
<i>Longshe Huo, Qingming Huang, Jianguo Xie</i>	
Watermarking (III)	
Semi-fragile Watermarking Based on Dither Modulation	721
<i>Jongweon Kim, Youngbae Byun, Jonguk Choi</i>	
An Adaptive Steganography for Index-Based Images Using Codeword Grouping	731
<i>Chin-Chen Chang, Piyu Tsai, Min-Hui Lin</i>	
DH-LZW: Lossless Data Hiding Method in LZW Compression	739
<i>Hiuk Jae Shim, Byeungwoo Jeon</i>	
Blind Image Data Hiding in the Wavelet Domain	747
<i>Mohsen Ashourian, Yo-Sung Ho</i>	
Image Watermarking Capacity Analysis Using Hopfield Neural Network	755
<i>Fan Zhang, Hongbin Zhang</i>	
Efficient Algorithms in Determining JPEG-Effective Watermark Coefficients	763
<i>Chih-Wei Tang, Hsueh-Ming Hang</i>	
A Fragile Watermarking Based on Knapsack Problem	771
<i>Hui Han, HongXun Yao, ShaoHui Liu, Yan Liu</i>	
Author Index	777

Generic Summarization Technology for Consumer Video

Masaru Sugano, Yasuyuki Nakajima, Hiromasa Yanagihara, and Akio Yoneyama

Multimedia Communications Laboratory, KDDI R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka, Saitama 356-8502, Japan

Abstract. This paper addresses automatic summarization technology for efficiently browsing video compressed by MPEG, which is a widely used for various consumer applications. By analyzing semantically important low- and mid-level features on compressed domain, the proposed method can universally summarize the MPEG video in the form of either *video skim* or *highlight*. Since all the summarization processes are performed completely on compressed domain, very fast summarization is achieved; only 20% of real-time playback in the case of SDTV. We develop the MPEG video summarization software which is capable of summarizing video and editing automatically created summary video.

1 Introduction

MPEG video has been widely used in consumer applications including personal video recording appliances such as DVD/HDD recorders, which enables consumers to store even more TV programs and enjoy time-shift capability than before. Also digital broadcasting and broadcasting system based on large capacity of storage devices such as TV-Anytime system [1] allow advanced browsing capabilities. For such large video archiving environments, video indexing, both structurally and semantically, plays a very important role in searching and browsing desired videos. Video summarization is one of the key elements in determining an efficient way to browse videos in storage media or over the network.

In this paper, we propose a generic summarization algorithm for efficiently browsing MPEG video. This algorithm analyzes the semantically important low- and mid-level features on compressed domain, automatically summarizes MPEG-1/-2 video at a low computational cost, and generates *video skims* or *highlights*. Video skim is a shortened version of an original while maintaining its context. Highlight, on the other hand, is an aggregation of important or exciting events obtained from TV sports programs. This useful technology can be applied to professional as well as consumer video applications. For professional use, we have developed automatic MPEG video summarization software which enables compressed domain summary indexing and MPEG video editing.

This paper is organized as follows. In Section 2, the previous video summarization works are overviewed. In Section 3, the proposed summarization method is described in detail. The experimental results are shown in Section 4. Also we introduce the MPEG video summarization software in Section 5.

2 Related Work

Video summarization is useful for obtaining a brief overview of an original video, and are roughly classified into two types, *video skim* and *highlight*. Many works on both abstraction techniques have been reported [2]-[9].

A video skim can be generated in the form of either a sequence of key frames [2][3] or a collection of key video clips [4]. Domain specific approaches have also been reported in movie trailers [5] and news sequence abstractions [6]. For example, a movie trailer is generated using face recognition, dialogue recognition, special event detection, etc. A more sophisticated approach uses a content model for audiovisual sequences based on various properties and relations of audio and visual segments [7]. On the other hand, a highlight may often be applied to TV sports programs. In this category, highlight generation algorithms have been proposed for baseball [8], American football [9], and so forth. The former utilizes only audio features, while the latter uses visual features such as textual overlay.

Though many summarization techniques have been proposed, few of them address generic and efficient summarization methods for MPEG video, which are widely used in consumer applications. Also most of the algorithms are domain specific or require a priori knowledge of contents such as baseball, or soccer. In the following section, we address a generic and efficient summarization algorithm that exploits low- and mid-level audio-visual features obtained on MPEG compressed domain.

3 Proposed Summarization Algorithm

3.1 Algorithm Overview

The proposed method achieves video clip based summarization and applies to both *video skim* and *highlight*. Basically, both audio and visual features of each shot are analyzed after shot-based segmentation, and then the summary segments are adaptively determined and assembled. In this subsection, the summarization algorithm is briefly overviewed. And in the following sub-sections from 3.2, features to be used and procedures are explained in more detail.

Video Skimming. In video skimming, generality and flexibility are two main advantages of our proposed methods. That is, this method can be applied to various kinds of videos such as movies, dramas, and documentaries. Furthermore, a user can specify arbitrary duration for video skim.

Our video skimming algorithm determines semantically significant shots from entire video based on scene analysis without a priori knowledge of content. Among the low-level features dedicated to semantic indexing, we especially focus on motion features, and classify scenes into dynamic or static using the *MotionActivity* descriptor defined in MPEG-7 [10]. This is based on the following assumptions: in dynamic scenes, more active shots are regarded as more

important, while in static scenes less active shots are regarded as more important. Typical examples are that in a documentary video, static or fixed shots rather than zooming or panning shots may often capture the important topics or objects, while in a sports video, dynamic or zooming shots rather than fixed shots may often capture the interesting events. These determined shots are then aggregated into a summary.

Highlight Extraction. Here, we focus on TV sports programs, and meaningful events are extracted using visual and audio features, from both *structured* sports, where the fixed shots are frequently appeared such as pitching scenes in baseball and rally scenes in tennis, and *non-structured* sports, where no fixed shots are included such as soccer.

Our method for extracting highlights is based on indentifying applause and/or detecting the recurrent (i.e. representative) shots using the MPEG-7 *ColorLayout* descriptor. That is, a summary can be extracted using the simple shot transition model revealed by recurrent shots detection. Typical examples are as follows: in TV sports videos, since the number of cameras and their positions are usually fixed and the obvious event boundaries are observed [11], similar shots are recurrently appeared, such as pitching scenes in a baseball game. These fixed shots or unvarying shots can be used for determining the event boundaries. For this purpose, we use the *ColorLayout* descriptor based histograms since this descriptor can be extracted from DC images created during shot-based segmentation, and finer discrimination can be achieved in terms of luminance and chrominance components.

3.2 Shot-Based Segmentation

Shot-based summarization is adopted in the proposed method, where a summary consists of a sequence of the key shots. Therefore, incoming MPEG video is at first segmented into shots by applying the shot detection algorithm [12]. Then the total number of shots NS is counted. In this stage, the shots with duration shorter than 3.5 seconds are discarded for human understandability [7].

3.3 Content Partition

This process is only required for video skimming to create user-defined duration of summary uniformly throughout an original video. After shot-based segmentation, incoming video is partitioned into equal-length partitions. The number of partitions NP is calculated as $NP = NS/2 \times SL/OL$, where OL , SL , and NS denote the original video length, the user-specified summary length, and the total number of shots, respectively. Partitioning allows uniform and flexible summarization from an entire video. For example, the last part of a movie which may include climax scenes can be intentionally omitted [5]. The content partition allows adaptive selection of the portions where summary shots will be extracted.

3.4 Audio Feature Analysis

Audio features are very important for summarization. This is because audio accompanied with video, such as narration in documentary videos, applause in sports videos, and events such as gunfire or explosions may often include semantically important video clips. Therefore, these features can be used for efficient and reliable summary creation.

Among several audio features, we simply employ audio energy calculated on MPEG audio subband domain. Since shots with silence or a very low audio level are not as significant as summaries in many cases, audio level analysis is conducted to exclude such shots using audio energy. On the other hand, shots with high audio level may be the candidate of summaries, for example, applause in sports video and some exciting events in action movies, etc. In this algorithm, audio energy is defined as subband-based weighted subband energy SBE_N in a shot N , derived from MPEG compressed audio [13]. In the following equation, sb_k denotes subband energy of the k -th subband ($k=1-32$).

$$SBE_N = 0.1 \times sb_1 + 0.2 \times \sum_{k=2}^7 sb_k + 0.7 \times \sum_{k=8}^{32} sb_k \quad (1)$$

3.5 Visual Features Analysis

The low-level visual features (e.g. still/static or dynamic, bright or dark, etc.) can be used for video analysis. In particular, motion characteristics are often used for scene classification [14]. To extract these characteristics properly, we use MPEG-7 visual descriptors [10] that can be easily obtained from compressed video. More specifically, the *MotionActivity* descriptor and the *ColorLayout* descriptor were selected as described in 3.1 because they can be easily calculated from MPEG coded parameters and provide good performances in similarity-based retrieval by low cost matching, which is used for further content clustering.

Motion Feature. *MotionActivity* represents the intuitive activity of a video segment, and can be used for content classification in terms of motion. Among the components of *MotionActivity*, the *Intensity* attribute is calculated from coded motion vectors in predicted frames. In MPEG-7, *Intensity* is defined as the standard deviation of motion vector magnitudes within a frame, and is expressed by an integer from 1 (lowest) to 5 (highest). The *MotionActivity* value for each shot is represented by the shot-averaged value. Our method utilizes a shot-averaged *Intensity* value of all the P-frames contained in each shot.

Although the key frame extraction using *MotionActivity* has been proposed [15], our method exploits it for estimating whether a shot is dynamic or static. Note that it depends on context whether a dynamic scene or a static scene is semantically significant. That is, if content is static such as a documentary video or a romantic movie, fixed or static shots are more important, while if content is dynamic such as a sports video or an action movie, shots with higher motion are semantically more important.

Color Feature. *ColorLayout* specifies spatial distribution of colors. It can be used for similar image/segment retrieval. When applying it to a video segment, a descriptor value for a representative frame within a segment can be used. Here, *ColorLayout* is extracted from a DC image generated at the shot segmentation stage [12] described in 3.2. Therefore, our method can greatly save the time for calculating a color feature and requires no decoding process. The DC image is first partitioned into 8×8 blocks. Then, averaged color is calculated in each block, and finally 8×8 DCT is performed on luminance and chrominance components, respectively. The value of *ColorLayout* is a set of DCT coefficients of these components. This color characteristic is mainly used for detecting a set of recurrent shots for structured sports video.

Here, we introduce a method for determining highly recurrent shots from an input video without a priori knowledge of content. This approach can be regarded as a similar shot retrieval without the predefined reference shot. That is, we employ histograms constructed using *ColorLayout* throughout the input video, and a set of the most frequent bins for each component is regarded as a reference *ColorLayout* value. After that, the distances D_N [13] between *ColorLayout* value of shot N and the above reference *ColorLayout* value are calculated, and if $D_N < Th_D$, the shot N is determined as a recurrent shot.

3.6 Adaptive Summary Determination and Shots Assembly

Video Skim Determination. The meaningful video skims should be decided according to content characteristics as described in 3.1 and 3.5. In our algorithm, at first, the summary shot candidates in each partition are determined according to content characteristics derived from the motion features shown in 3.5. That is, *Intensity of MotionActivity* is analyzed and each partition is determined as being either static or dynamic, and summary shots are determined by adaptive thresholding. For example, when a partition-averaged *Intensity* I_P is below Th_I ($=2.5$), that partition is regarded as static, and then shots with shot-averaged *Intensity* I_S below I_P and audio energy *SBE* above Th_S (a threshold for silence) are extracted from the partition in ascending order until the sum of the extracted shot lengths exceeds the content-averaged shot length AL ($=OL/NS$), and vice versa. This process results in video skimming with user-defined duration SL ($=2 \times NP \times OL/NS$). Finally, the extracted shots from all partitions are assembled to construct a video skim.

Highlight Determination. Highlight is determined based on the context of audio only for non-structured sports, or audio and video for structured sports. For context of audio, the following assumption can be applied [13]; if a highlight exists, loud applause from audience occurs just after the exciting events. Therefore, the audio energy *SBE* of each shot are evaluated, and if a peak *SBE* exists, a highlight is determined prior to the corresponding shots. In addition, as for structured sports, the frequently appeared shots are determined at first, and if the corresponding shots and/or subsequent 1-2 shots have a peak or sufficiently large *SBE*, a sequence of shots can be determined as a highlight segment.

Table 1. Video skimming results

Video	Original length	Summary length	# paragraphs	# matching	Recall
Movie 1	03:06:00	00:17:14	22	17	77.3%
Movie 2	01:55:00	00:11:19	20	14	70.0%
Drama 1	00:57:11	00:04:56	9	6	66.7%
Drama 2	00:59:00	00:06:18	6	5	83.3%

4 Experimental Results

Experiments were performed to verify the proposed summarization algorithm, in terms of video skimming and highlight extraction. In the experiments, MPEG-1 videos (SIF) were used and extracted summary segments were assembled into a new MPEG-1 summary video using the MPEG editing SDK [16].

4.1 Video Skimming

In evaluating video skimming algorithm, two movies (*“Titanic”* and *“Raiders of The Lost Arc”*) and two TV dramas (a Japanese drama *“Big White Tower”* and a Korean drama *“Winter Sonata”*) were used. From these videos, video skims with 10% of the original length were generated. As for video skimming, it is very difficult to objectively evaluate the results since the valid evaluation criteria have not been established. Therefore, in this experiment, we compared the contexts between skimmed videos and textual abstracts available on one or more web sites (official/unofficial). That is, the textual abstracts are first splitted into arbitrary number of paragraphs, and if skimmed video includes the contexts described in a certain paragraph, then the corresponding paragraph is determined to match. Table 1 shows the original/the summarized video length, the number of splitted paragraphs, the number of paragraphs matched, and recall.

As shown in Table 1, our proposed method successfully summarizes video while maintaining its context. Also adaptive determination shown in 3.6 is effective since Movie 1 (*“Titanic”*), which includes both action and romantic scenes, is summarized at high recall rate, for example. Such video skims provide users a preview on the contents and help remind users of its outline in short time, which is useful for both consumer and professional video applications. All the processes on the normal Windows PC, excluding compressed domain editing, require only about 80MHz and 400MHz or only 4% and 20% of real-time playback for MPEG-1 video (SIF) and MPEG-2 video (SDTV), respectively. These are reasonable processing load and speed for consumer video browsing/recording devices.

4.2 Highlight Extraction

Structured Sports. To evaluate the recurrent shots detection algorithm described in 3.5, we used two sumo videos and one tennis video, where the recurrent shots themselves are significant events (matches in sumo, rallies in tennis). Table

Table 2. Recurrent shots detection results

Video	Detected	Correct	Precision	Recall
Sumo 1 (105 min)	24	20	83.3%	95.2%
Sumo 2 (48 min)	11	9	81.8%	100%
Tennis 1 (30 min)	54	53	98.1%	98.1%

Table 3. Soccer highlights results

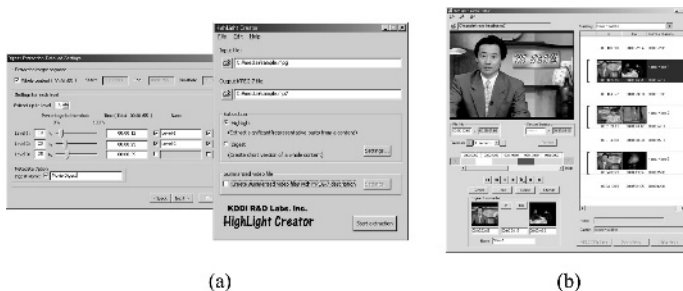
Video (#G/#S)	Detected	Correct (#G/#S)	Precision	Recall
Soccer 1 (4/10)	16	14 (4/10)	87.5%	100%
Soccer 2 (3/9)	16	12 (3/9)	75.0%	100%
Soccer 3 (2/9)	11	10 (2/8)	90.9%	90.9%

2 shows the number of detected recurrent shots. As seen in the table, the recurrent shots detection using *ColorLayout* results in high recall. By combining the audio feature described in 3.6, only exciting matches and rallies are extracted.

Non-structured Sports. As for non-structured sports video, we used three soccer videos of 45-minutes duration. The threshold of audio energy for applause was determined using a different 45-minutes soccer video. Here, the number of detected goals (G) and shoots (S) was evaluated. Table 3 shows the results, which indicate that our proposed algorithm effectively extracts soccer highlights.

5 MPEG Summarization Software

Figure 1 shows the GUI of *Highlight Creator*, the automatic MPEG summarization software which we have developed for video productions based on the above mentioned methods. This software is capable of summarizing MPEG video (*Logger*, Figure 1 (a)) and editing automatically created summary video (*Editor*, Figure 1 (b)). Also the summarized video can be converted into various formats for Internet streaming (WindowsMedia, RealMedia) or 3G mobile phone (3GPP2).

**Fig. 1.** MPEG Summarization Software “*Highlight Creator*”

6 Conclusions

This paper proposed an efficient and computationally low-cost MPEG video summarization. By using semantically significant characteristics on compressed domain, the meaningful summaries of video skims and highlights were successfully extracted. The authors thank Dr. T. Asami, Dr. S. Matsumoto and Dr. M. Wada for their continuous support. The authors also would like to thank Mr. Y. Nagata and Mr. M. Sawaguchi for their efforts in testing our algorithm.

References

1. ETSI TS 102 822-3-1: Broadcast and on-line services: Search, select, and rightful use of content on personal storage systems (“TV-Anytime Phase 1”); Part 3: Metadata; Sub-part 1: Metadata schemas. June 2004.
2. A. Hanjalic and H. Zhang: An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, pp.1280-1289, December 1999.
3. A.M. Ferman, A.M. Tekalp: Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Trans. on Multimedia*, Vol. 5, No. 2, pp. 244-256, June 2003.
4. Y. Gong and X. Liu: Video summarization with minimal visual content redundancies. *IEEE ICIP 2001*, Vol. 3, pp. 362-365, October 2001.
5. R. Lienhart, S. Pfeiffer, and W. Effelsberg: Video abstracting. *Communications of the ACM*, Vol.40, No.12, pp. 55-62, December 1997.
6. Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray: Automated generation of news content hierarchy by integrating audio, video, and text information. *IEEE ICASSP 99*, Vol. 6, pp. 3025-3028, March 1999.
7. J. Saarela and B. Merialdo: Using content models to build audio-video summaries. *SPIE Conference on Storage and Retrieval for Image and Video Database VII*, Vol. 3656, pp. 338-347, January 1999.
8. Y. Rui, A. Gupta, and A. Acero: Automatically extracting highlights for TV baseball programs. *ACM Multimedia 2000*, pp. 105-115, October 2000.
9. N. Babaguchi, Y. Kawai, and T.Kitahashi: Generation of personalized abstract of sports video. *IEEE ICME 2001*, pp. 800-803, August 2001.
10. ISO/IEC 15938-3: Information Technology – Multimedia Content Description Interface – Part 3: Visual. July 2002.
11. S.-F. Chang and H. Sundaram: Structural and semantic analysis of video. *IEEE ICME 2000*, pp. 687-690, July 2000.
12. Y. Nakajima, et al.: Universal scene change detection on MPEG coded data domain. *IEEE VCIP 97*, Vol. 3024, pp. 992-1003, February 1997.
13. M. Sugano, Y. Nakajima, and H. Yanagihara: MPEG content summarization based on compressed domain feature analysis. *SPIE Conference on Internet Multimedia Management Systems IV*, Vol. 5242, pp. 280-288, September 2003.
14. Y. Wang, et al.: Multimedia content classification using motion and audio information. *IEEE ISCS 97*, Vol. 2, pp. 1488-1491, June 1997.
15. A. Divakaran, et al.: Motion activity-based extraction of key-frames from video shots. *IEEE ICIP 2002*, Vol. I, pp. 932-935, September 2002.
16. MP-Factory: <http://avs.kddilabs.jp/mpeg/mpfs40/indexe.html>. August 2004.

Movie-in-a-Minute: Automatically Generated Video Previews

Mauro Barbieri¹, Nevenka Dimitrova², and Lalitha Agnihotri²

¹ Philips Research The Netherlands, Prof. Holstlaan 4, 5656AA Eindhoven

² Philips Research USA, 345 Scarborough Road, Briarcliff Manor, NY 10510
{mauro.barbieri, nevenka.dimitrova, lalitha.agnihotri}@philips.com

Abstract. *Movie-in-a-minute* is a summarization method that enables quick browsing and access to hundreds of hours of stored video programs. A *movie-in-a-minute* is a short video sequence composed of automatically selected portions of the original video that aims at conveying key aspects of a program and its story in an efficient but entertaining way. In this paper we discuss an approach to generating *movie-in-a-minute* summaries using film grammar rules to guide the selection of video segments that are indexed using automatically computed signal-level features.

1 Introduction

Summarization has become a highly necessary tool in browsing and searching home video collections and produced video archives, saving users' time and offering great control and overview. Various types of summarization methods have been offered in the literature: visual table of contents, skimming, multimedia summaries [1]. Also, various domains have been explored such as structured video summarization for news, music videos, and sports. On the other hand, the Holy Grail remains to be narrative video summarization, which includes methods for summarizing narrative content such as movies, documentaries and home videos. In this paper we present *Movie-In-A-MInute (Miami)*, also known as "video preview" or "video thumbnail." *Miami* is a short video sequence dynamically composed of selected portions from the original video. It aims at conveying key aspects of a program and its story with an array of important images and segments. *Miami* videos help users selecting programs (instead of zapping channels), deleting, downloading or simply recalling watched programs. These features are indispensable for large video archives, such as personal video recorders and home network systems [2].

While video summaries aim at conveying all the information of the original content in shorter and more efficient versions, *Miami* videos aim only at giving users clues for selecting programs. Therefore a *Miami* video does not need to be comprehensive, or to include all highlights. Video trailers are somewhat similar to *Miami* videos although they purely aim at teasing consumers, attracting their attention and creating expectations that can or cannot be met by consuming the real content.

In this paper we will first discuss related work to video summarization in section 2. We will detail the requirements for *Miami* video in section 3. Then, in section 4 we will introduce a formal model within an optimization framework that translates the requirements into constraints. Implementation and results of this model will be given in section 5. Section 6 will introduce the need for personalization of video summaries, and section 7 will conclude the paper.

2 Related Work

In the recent literature various video summarization methods have been introduced: *video skim*, *highlights*, and *multimedia summaries*.

Video skim is a temporally condensed form of the video stream that preferably preserves the most important information. It is a set of short video sequences composed of automatically selected portions of the original video. A method for generating visual skims based on scene analysis and using the grammar of film language is presented by Sundaram et al. [6]. Ma et al. [4] proposed an attention model that includes visual, audio, and text modalities for video summarization. With respect to these summarization methods, *Miami* videos are not meant to convey all the information of the original content but aim at including only key aspects of a program to allow users to quickly see what it is about and make a selection.

Video highlights is a form of summary that aims at including the most important events in the video. Various methods have been introduced for extracting highlights from specific genres of sports programs: goals in soccer video, hits in tennis video or pitching in baseball [7], important events in car racing video [5], and others.

Multimedia video summary is a collection of audio, visual, and text segments that preserve the essence and the structure of the underlying video (e.g. pictorial summary, story boards, surface summary). A multimedia video summary of audio-video presentations is presented in [3]. The summarization system uses slide-transitions in video, pitch in audio and user interactions with presentations in order to generate a multimedia summary.

3 Requirements

The automatic creation of a *Miami* video can be formulated as the problem of selecting the best set of segments of a given duration of the original program that satisfies a certain list of requirements.

Two different approaches can be followed for the design of an algorithm that generates *Miami* videos: the *machine learning* approach and the *knowledge-explicit* approach. In the *machine learning* approach a statistical classifier is trained with positive and negative examples, selected by humans, with the aim of generalizing the common underlying properties of the examples in such a way that the classifier learns to distinguish between “good” previews and “bad” previews. This approach has the advantages of being generic and potentially

reusable for all types of video and video previews. However, in practice the main problem is the amount of proper positive and negative examples required to train a classifier that can achieve a proper level of generalization. A simpler problem that could be tackled using machine learning is deciding for each segment of the original program whether or not it is suitable for being included in a preview. Although appealing, this method neglects to consider that a good preview is not simply a preview formed by including “good” segments. The meaning conveyed by a video lies largely in the relationships and the temporal order of the segments of which it is composed.

In the *knowledge-explicit* approach, the designer embeds in the algorithm the knowledge on how to make a video preview in the form of requirements and constraints that drive the search for the best subset of the original program and the composition of the video preview. Machine learning can then still be used to fine-tune the parameters of the model in an objective and systematic way. Based on a study of cinematic production rules we have developed our own knowledge-explicit approach.

To allow fast and convenient content selection, a video preview should meet more than thirty requirements that have been collected by analyzing related literature on video summarization (i.e. [4][8]) and film production (i.e. [9][10]) and interviewing a restricted number of “expert” users. The requirements can be grouped into seven categories: *duration*, *continuity*, *priority*, *uniqueness*, *exclusion*, *structural* and *temporal order*:

- *Duration* requirements deal with the durations of the preview and of its subparts. Each segment chosen for the preview has a minimum time required for comprehension depending on its type, complexity, and generically speaking, amount of information it conveys.
- *Continuity* requirements necessitate that a video preview should be as continuous as possible; users will not appreciate a preview with many abrupt “jumps.”
- *Priority* requirements indicate which content should be included in the preview to convey as much information on the program as possible in the shortest amount of time. Examples are: including sequences with close-ups of the main actors as well as action segments and dialogues giving clues on the story line.
- *Uniqueness* requirements aim at maximizing the efficiency of the preview by minimizing redundancy.
- *Exclusion* requirements indicate which content should not be included in the preview. For example a preview of a recorded broadcast program should not include any commercial advertisement. Additionally, in order not to spoil the plot of the program and allow users to later view the content in its entirety, the video preview should not disclose the end of the story.
- *Structural* requirements dictate rules that pay attention to the structural properties of video. For example, in order to provide a good overview, a video preview should cover uniformly the entire program and mimic its original mood and tempo.

- *Temporal order* requirements concern the temporal order of the sequences included in the preview. In this category, users have indicated conflicting requirements. Keeping the original order certainly helps users to understand the story line given the few clues provided by the preview. On the other hand, changing the order prevents revealing too much of the story line in case users want to later view the entire content. The choice of which requirement to follow can be left to the final user of the system.

The requirements are formalized in computable constraints in order to be used to select a subset of segments from the original program that is admissible for being a video preview. At the same time, to create a good preview it is necessary to compare admissible sets of segments to select “the best” set. The comparison is based on a function that numerically estimates the value of a preview, the *importance score*. Given a computational model of the constraints and an importance score function to maximize, the problem of automatic generation of video previews is a constrained optimization problem and its solution can be found with known methods (e.g. constraint logic programming, local search techniques [19]).

4 Formal Model

In this section the previously listed requirements are mapped to a part of an objective function or to a constraint that the *Miami* video should fulfill.

In formalizing the problem of automatic preview generation, the original program can be seen as a finite sequence of successive video segments, with synchronized audio and subtitle $V = v_1 \cdots v_M$ where v_i is the i -th video segment in the original program. M depends on the original program duration and on the actual segmentation. The desired video preview can be represented as a finite sequence of successive positions that can be taken by any video segment belonging to the original program: $S = s_1 \cdots s_N$ where s_j is the j -th position in the preview. N depends on the duration of the preview that is fixed by the user to a certain amount D : $\sum_{j=1}^N duration(s_j) \leq D$. The duration of each video segment should not be shorter than a certain minimum amount, to be understandable out of its original context and, at the same time, it should not be longer than a certain maximum value in order not to give away too many details of the story. This can be formalized by requiring: $d_{min} \leq duration(s_j) \leq d_{max}$.

The objective function whose absolute maximum denotes the best preview, has the following structure:

$$eval(S) = \alpha \cdot \pi(S) - \beta \cdot \rho(S) + \gamma \cdot \eta(S) + \delta \cdot \omega(S) . \quad (1)$$

In (1) the priority requirements are taken into consideration in $\pi(S)$ that is defined as:

$$\pi(S) = \sum_{j=1}^N \pi(s_j) \quad (2)$$

where $\pi(s_j)$ is the priority score of segment s_j and it is defined as follows:

$$\pi(s_j) = \mathbf{w} \cdot \mathbf{A}(s_j) \quad (3)$$

in which \mathbf{w} is a vector of weighting factors and $\mathbf{A}(s_j)$ is a column vector of attributes associated to segment s_j in the range $[0 \cdots 1]$. These attributes are computed by applying several low- and mid-level content analysis algorithms such as: computation of contrast, audio loudness, detection of action [11], faces [12], dialogues [13], music/speech/noise/silence [14], and camera motion [15]. The relative importance of the various attributes can be linearly tuned using the weighting factors \mathbf{w} .

The term $\rho(S)$ in equation (1) estimates the degree of redundancy of the preview that in our case has a negative sign, which means we promote uniqueness and penalize redundancy. It is defined as a linear combination of visual and textual redundancy:

$$\rho(S) = \beta_1 \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sigma_v(s_i, s_j) + \beta_2 \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sigma_t(s_i, s_j) \quad (4)$$

where $\sigma_v(s_i, s_j)$ represents the visual similarity of segments s_i and s_j and is computed based on automatically extracted visual features. Textual redundancy is measured by extracting keywords in the close captions or in the speech transcript, $K(s_i)$, and by counting the number of times they are repeated in the preview segments:

$$\sigma_t(s_i, s_j) = |K(s_i) \cap K(s_j)| \quad (5)$$

The continuity requirements can be taken into account by considering the shots as elementary segments constituting the program. The shot boundaries can be found by performing shot cut detection. Additionally, sentences should be included entirely and not be abruptly cut while subtitles should be displayed for a sufficient amount of time to be read. If we represent the synchronized audio stream A as $A = a_1 \cdots a_Q$ (a_j being the j -th audio segment and Q the number of audio segments), the synchronized subtitles C as $C = c_1 \cdots c_R$ (c_k being the k -th subtitle and R the total number of subtitles), and we indicate with b_s , e_s , and Δ_s respectively the start time, the end time and the time-span of the video, audio or subtitle segment s , the continuity requirement of including complete audio segments can be formalized with the following constraints (the same applies to subtitles):

$$\forall s \in S (\forall a : e_a \in \Delta_s, b_a \in \Delta_s) \wedge (\forall a : b_a \in \Delta_s, e_a \in \Delta_s) \quad (6)$$

The requirement of not including commercial blocks can be easily fulfilled by removing the segments indicated by our commercial block detector [16]. Additionally, we take into consideration the requirement of not disclosing the end of the program by discarding a certain percentage¹ of segments at the end (e.g. 10%).

¹ A statistically sound percentage can be found by identifying for a large set of programs at which point the end is disclosed.

The structural requirement of uniform coverage of the whole program can be fulfilled by considering a segmentation of the program into L different scenes (U_j , $j = [1 \cdots L]$) and maximizing $\eta(S)$ in equation (1) that is the product of the relative durations of the selected segments belonging to each scene:

$$\eta(S) = \sqrt[L]{\prod_{j=1}^L \frac{\sum_{s \in U_j} \text{duration}(s)}{\text{duration}(U_j)}}. \quad (7)$$

Scene boundaries are computed using a time-constrained clustering procedure similar to the one described in [17].

Temporal order requirements different from the original order are taken into consideration in the term $\omega(S)$ of equation (1). For example, to generate a preview having all the action segments at the end, $\omega(S)$ is defined as follows:

$$\omega(S) = \sum_{i=1}^N i \cdot \text{action}(s_i) \quad (8)$$

where $\text{action}(s_i)$ indicates whether segment s_i is classified as action segment. The original order constraint is implicitly solved during optimization by keeping the video segments chronologically ordered.

The weights α , β , γ , and δ in (1) allow personalizing the generation of the preview (as discussed in section 6) by changing the relative impact of the different types of requirements on the value of the objective function.

5 Implementation and Results

The generation of a *Miami* video can be divided into four main steps (figure 1):

1. Audio and video feature extraction.
2. Audio and video segmentation and classification.
3. Segments selection.
4. Preview composition.

5.1 Audio and Video Feature Extraction

Various algorithms are applied to the audio and video signals to extract the features required by the next steps and for computing the priority score according to equation (3) after normalization over the entire video. Video features include low-level attributes such as contrast, color distribution, motion activity and mid-level attributes such as face location and size, camera motion, etc. Audio features (see [14] for a detailed description) include RMS value, spectral centroid, bandwidth, zero-crossing rate, MFCC, etc.

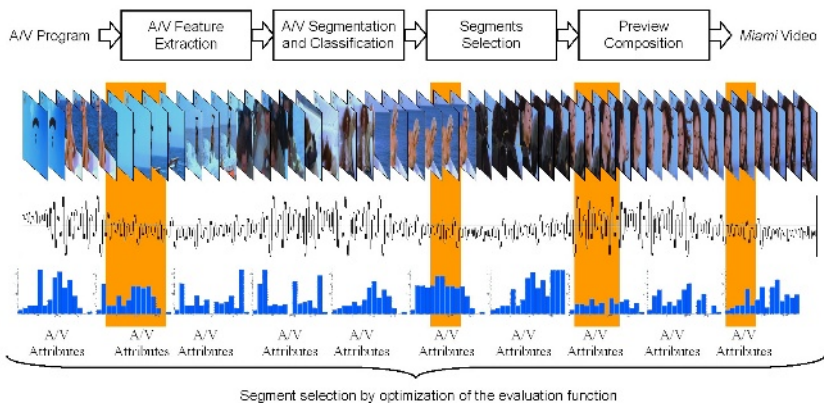


Fig. 1. Steps for the generation of a *Miami* video

5.2 Audio and Video Segmentation and Classification

This step can be divided into five sub-steps:

2a. *Shot segmentation and clustering*: a standard shot cut detection algorithm [18] is applied to divide the video stream into continuous shots. Time-constrained clustering [17] is then applied to group together visually consistent shots that are not far apart in time.

2b. *Audio classification*: the synchronized audio stream is classified into coherent audio classes such as silence, speech, music, noise, etc. Changes in the audio class indicate the audio segment boundaries used to verify constraint (6).

2c. *Micro-segmentation*: segments exceeding the maximum duration after the shot segmentation are further divided into sub-segments with durations bigger than d_{min} and with boundaries possibly aligned with content-based clues such as: a change in the audio class, appearance or disappearance of a detected face, a change in camera motion or object motion. The micro-segmentation step can be easily formalized as an integer linear programming problem and solved with standard methods (e.g. simplex method).

2d. *Segment compensation*: successive segments violating constraint (6) are merged until the continuity requirement is fulfilled without violating the maximum segment duration d_{max} . When this is not possible, the segmentation induced by the audio classifier is used as primary instead of the shot-based one.

2e. *Pre-filtering*: commercial detection [16] is performed over the entire video and the detected commercials are discarded from the set of segments available for the generation of the *Miami* video. In order not to disclose the end of the program, an extra 10% of segments is removed from the end.

5.3 Segments Selection

The segments selection step consists of searching the best set of segments that maximize the objective function (1) in the space of all possible previews. The

space of all possible previews can be explored using a *local search* method such as simulated annealing or a genetic algorithm [19] because at this point each requirement or constraint has either been solved in the previous steps or it is mapped to a priority or penalty score term in the objective function (1). However, considering that the actual usage of the *Miami* video is to provide a rough overview of the content of a program, the goal of finding the absolute maximum of (1) can be relaxed to finding a good approximation, a preview with a reasonably high value of $eval(S)$.

We have implemented a heuristic search strategy that iteratively improves an initial set of selected segments. The starting set is constructed by selecting for each scene the segment with the highest priority score $\pi(s_j)$ that generates the minimum redundancy $\rho(S)$. At every iteration the first segments of each scene that improve the objective function are added to the set. The algorithm stops after a certain fixed number of iterations or if $eval(S)$ cannot be significantly improved. The solution is not optimal but usually good enough for the typical *Miami* video usage.

5.4 Preview Composition

The last step consists of the actual composition of the preview by fusing the selected segments into one continuous audiovisual stream. Abrupt audio and video transitions between segments are smoothed using dissolve effects.

5.5 Prototype Implementation and Results

A prototype of *Miami* video has been implemented in *C++* and *Java* for the generation of previews of recorded broadcast programs in MPEG-2 format. The generation of a *Miami* video on a state-of-the-art personal computer requires no longer than the actual program duration. Most of the CPU time is used for video decoding and content analysis algorithms; the segments selection step requires only a fraction of the total running time.

In preliminary tests the system has been manually tuned and tested with a large set of narrative programs such as feature films and documentaries. The typical duration of a *Miami* video for a two-hours-long feature film is usually set to 60 or 90 seconds.

When we presented the generated previews to users we always received very positive feedback. A series of formal tests with users is currently being designed to evaluate the quality of the preview by judging fluency, clues and coherency of the story line, and to evaluate how the feature helps users in selecting video content from a large collection.

6 Need for Personalization

As content availability continues to grow, personalized summaries become important. For example a movie mainly containing action scenes could also have

a poignant love story embedded in it. Persons who particularly like love stories might like previews highlighting these love story elements. Users will require summaries to be personalized so that they can choose the movie they like to watch and not miss out on a movie because the preview did not include sections that might appeal to them more.

Any of the requirements of duration, continuity, priority, uniqueness, exclusion, structural, and temporal order, that were presented in section 3, can be subject to personalization. For example, a user might desire to see more of the introduction segments, which will then affect the structural requirements. The priority requirements can also be based on a user profile: for a person who prefers “dark,” “silent” scenes, we should include those as opposed to “bright,” “dialog” scenes.

So far the user preferences on summarization have not been fully explored by the research community. An exploration panel of experts and users [20] on issues of multimedia summarization indicated that summaries should be personalized.

As with any personalization, the problem is twofold: to have an extensive good profile that reflects the user’s needs and to have an accurate model for performing the computational matching of the user profile to the video analysis features. The challenge here is to ask the “right” questions in order to generate this user profile. One approach is to pose this as a problem of learning from examples where users would be shown many previews and would need to select the one that appeals to them the most. Once the system is trained to the type of previews that a user likes, the different weights that were presented in section 4 can be worked out in order to generate personalized previews.

7 Conclusions

Producing movie trailers in the production world is an art in itself. On the other hand, previews, or as we call them *movie-in-a-minute (Miami)* videos are not available for all the different types of narrative programs and home videos. Moreover, people with various tastes would like to see personalized previews. In this paper we introduced a knowledge-based computational framework for generating *Miami* videos that includes: audio and video feature extraction, audio and video segmentation and classification, segments selection, and preview composition.

The framework has been implemented and manually tuned for narrative type of content such as feature films and documentaries. In preliminary tests, users were impressed by the quality achieved by the algorithm.

Future work will include a formal method to fine-tune the model parameters and evaluation of the results based on user tests.

References

1. Barbieri M., Agnihotri L., Dimitrova N., “Video Summarization: Methods and Landscape,” Proc. of SPIE Int. Conf. on Internet Multimedia Management Systems IV, ITCOM 2003, Orlando, USA, 7-11 September 2003

2. Paulussen I., Barbieri M., Mekenkamp G., "The SPATION Project: Embedding Content Analysis in Consumer Electronics Networks," Proc. of the Third International Workshop on Content-Based Multimedia Indexing, CBMI 2003, Rennes, France, September 2003
3. He L., Sanocki E., Gupta A., Grudin J., "Auto-Summarization of Audio-Video Presentations," Proc. of ACM Multimedia 1999, Orlando, USA, November 1999
4. Ma Y.-F., Lu L., Zhang H.-J., Li M., "A User Attention Model for Video Summarization," Proc. of ACM Multimedia 2002, Juan Les Pin, December 1-5, 2002
5. Petkovic M., Mihajlovic V., Jonker W., "Multi-modal extraction of highlights from TV formula 1 programs," Proc. of IEEE Int. Conf. on Multimedia and Expo, ICME 2002, Lausanne, Switzerland, 2002
6. Sundaram H., Xie L., Chang S.-F., "A Utility Framework for the Automatic Generation of Audio-Visual Skims," Proc. of ACM Multimedia 2002, Juan Les Pin, December 1-5, 2002
7. Zhong D., Kumar R., Chang S.-F., "Demonstrations: Real-time personalized sports video filtering and summarization," Proc. of ACM Multimedia 2001, October 2001
8. Pfeiffer S., Lienhart R., Fischer S., Effelsberg W., "Abstracting Digital Movies Automatically," Journal of Visual Communication and Image Representation, Vol. 7., No. 4, December 1996
9. J. V. Mascelli, "The Five C's of Cinematography - Motion Pictures Filming Techniques," Silman-James Press, Los Angeles, CA, USA, 1965
10. H. Zettl, "Sight Sound Motion - Applied Media Aesthetics," Third Edition, Belmont, CA, USA, Wadsworth Publishing Co., 2001
11. Peker K. A., Divakaran A., Papatthomas T. V., "Automatic Measurement of Intensity of Motion Activity of Video Segments," Proc. of SPIE Storage and Retrieval for Media Databases, Vol. 4315, 2001
12. Abdel-Mottaleb M., Elgammal A. "Face Detection in Complex Environments from Color Images," Proc. of Int. Conf on Image Processing, ICIP 1999
13. Sundaram H., Chang S.-F., "Determining computable scenes in films and their structures using audio-visual memory models," Proc. of ACM Multimedia 2000, Marina del Rey, USA, 2000
14. McKinney M., Breebaart J., "Features for Audio and Music Classification," Proc. of the 4th Int. Symposium On Music Inform. Retrieval, ISMIR 2003, Washington DC, USA, October 2003
15. Tan Y.-P., Saur D., Kulkarni S., Ramadge P., "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 1, February 2000
16. Schaffer D., Agnihotri L., Dimitrova N., McGee T., Jeannin S., "Improving Digital Video Commercial Detectors with Genetic Algorithms," Proc. of the Genetic and Evolutionary Computation Conference 2002, New York, July 2002
17. Boreczky J., Girgensohn A., Golovchinsky G., Uchihashi S., "An Interactive Comic Book Presentation for Exploring Video," Proc. of ACM CHI 2000, Vol. 2, No. 1, The Hague, The Netherlands, April 2000
18. Lienhart R., "Comparison of Automatic Shot Boundary Detection Algorithms," Proc. of Storage and Retrieval for Image and Video Databases VII, Vol. 3656, San Jose, USA, January 1999
19. Aarts E.H.L., Lenstra J.K., "Local Search in Combinatorial Optimization," John Wiley & Sons, Chichester, England, 1997
20. Agnihotri L., Dimitrova N., Kender J.R., Zimmerman J., "Study on Requirement Specifications for Personalized Multimedia Summarization," Proc. of IEEE Int. Conf. on Multimedia and Expo, ICME 2003, Baltimore, USA, July 2003

Automatic Sports Highlights Extraction with Content Augmentation

Kongwah Wan¹, Jinjun Wang^{2,1}, Changsheng Xu¹, and Qi Tian¹

¹ Institute for Infocomm Research,

21 Heng Mui Keng Terrace, Singapore 119613

² Nanyang Technological University, SCE, Singapore 637598

{kongwah, stuwj2, xucs, tian}@i2r.a-star.edu.sg

Abstract. We describe novel methods to automatically augment content into video highlights detected from soccer and tennis video. First, we extract generic and domain-specific features from the video to isolate key audio-visual events that we have empirically found to correlate well with the ground-truth highlights. Next, based on a set of heuristics-driven rules to minimize view disruption, spatial regions in the image frames of these video highlight segments are segmented for content augmentation. Preliminary trials from subjective viewing indicate a high level of acceptance for the content insertions.

1 Introduction

With the world-wide growth of consumer devices such as mobile phones and home set-top-boxes (STB), content providers are looking for sustainable revenue streams from innovative video applications in their distribution networks. With its global appeal, sports video is widely seen as a key driver content to launch applications such as interactive TV. Significant research effort has also been devoted to the automatic extraction of sports highlights in the past few years [1]. In particular, interesting results from our recent work in [2] points to the potential for a secondary market for game viewer-ship on mobile devices, and that replay selection and generation is no longer the exclusive purview of the game broadcasters. Issues naturally arise as to how the business case can be further enhanced. We note that the traditional model of 30-sec advertising-run for broadcast TV may not work well on mobile platform for obvious reasons of bandwidth cost and duration ratio. In contrast, in-program content augmentation appears to be more suitable. However, existing techniques such as [3] are generally hardware-driven and expensive. Our intention in this paper is to explore alternative techniques for content insertion (used interchangeably in this paper with content augmentation) and how they can be integrated with automatic sports highlights extraction for consumer video applications. We organize the rest of the paper as follow. We first describe automatic highlight extraction for soccer and tennis in Section 2. Then we provide an overview of our content insertion techniques in Section 3. Experimental results are detailed in Section 4 before concluding on some future work in Section 5.

2 Automatic Sports Highlights

As most sports events are played in constrained settings and telecasted with a small number of cameras, the limited view categories and their repetition is intuitively amenable for techniques such as shot type classification [4]. Structures in soccer [5] and tennis [6] have been capitalized for play-break classification and high level retrieval. For low end devices, compute-efficient techniques have been developed in [7,8] using a small set of generic audio-only features. However, it should be pointed out that most of these literature address the problem of event *isolation*, ie, its start-point, rather than its boundary end-points, which is more difficult as it requires a proper enclosure of the pre-event context and post-event response. In what follows, we give a brief overview on our key recent contributions [2,8,9] in event boundary detection.

2.1 Audio End-Points (Soccer): Excited Commentary

The difficulty in analyzing soccer audio is compounded by the large spectral variation in mixing commentary speech, field audio and noise. Figure 1 (left) replicates the system in [8], where we approach the problem as locating the “dominant speech” portions in the audio, which we model as a band-limited pulse train in the frequency domain. A novel Composite Fourier Transform (CFT) is then used to isolate the excited commentary: a first FT of the signal is applied, and the resulting magnitude spectrum is treated as another time domain signal input to a second FT. This differs from the widely-used cepstrum method, which applies an inverse FT to the log-spectrum instead. To calculate the end-points, voiced segments based on MFCC features are input to a robust pitch tracker to mark the rise and fall of excited commentary.

2.2 Visual End-Points (Soccer): Shot-Cut Rate, Field Positions, Goal-Mouth, etc.

Figure 1 (right) shows the 2 additional visual features used in [9] to compute highlight boundary end-points: shot-change rate and goal-mouth appearances. The former intuitively capture the production rule that after a significant event, views from multiple cameras are used to reproduce the action from alternative view angles or to survey the on-field emotions. In particular, computing shot cuts on video from panning cameras

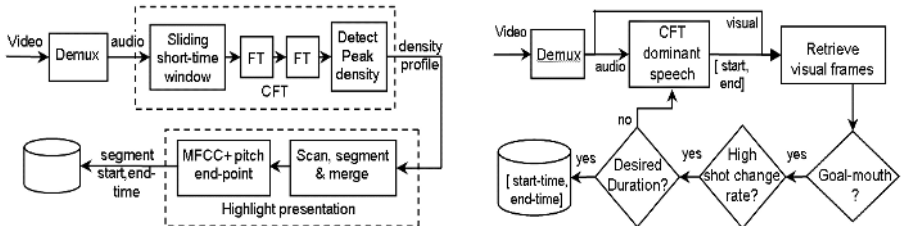


Fig. 1. Left: Composite Fourier Transform-based highlight detection on soccer audio; Right: Inclusion of visual features for end-point calculation

tracking the paths of celebrating players usually result in rapid cut sequences of short duration. Together with a reliable detection of goal-mouth appearance, the number of stray outputs from an audio-only (CFT) approach is effectively reduced.

Our other recent paper in [2] further describes a novel approach to generate soccer replay segments from *only* the main panoramic video camera. When used in a multiple-camera-setup, the end-points can be used as time-stamp markers to collate the other video streams, and further analysis can be made as to which is a better video feed to go on air. Not only can these segments be replay candidates for decision by the broadcast TV director, they can also be distributed to a secondary channel of viewers on, say, mobile platform. A mid-level representation framework is used to create audio-visual keywords from the low level features: (Visual) F_1 : Active Play Position, F_2 : Ball trajectory, F_3 : Goal mouth location, F_4 : Motion activity; (Audio) F_5 : Whistling, Acclaim. Three types of events are defined: Attack, Foul and Others. Three SVM classifiers (Gaussian kernel) are trained on an empirically derived set of mid-level keywords: Attack: F_1, F_2, F_3 and F_5 ; Foul: F_1, F_4 and F_5 ; Other: F_1 and F_4 .

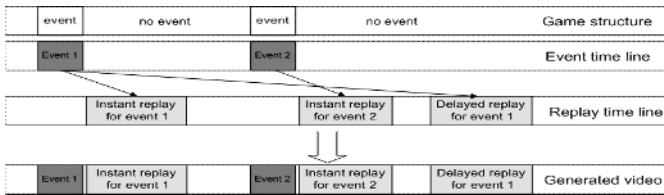


Fig. 2. Rules for replay insertion

To compute replay boundary, a search algorithm is applied to search backward and forward from the moment of event occurrence. The backward search checks whether the play position keyword F_1 has changed from $t_s - D_1$ to $t_s - D_2$, where t_s is the event moment starting time and D_1, D_2 are the minimal and maximal offset threshold respectively. A similar forward search is applied to detect the event end-time. These boundaries are input to the replay generation module (Figure 2). For any event segmented, the system attempts an *instant* replay by examining whether it can be inserted at the following “No-event” slot (instant replays for event-1 and 2 in row-2 and 3). Otherwise, it checks for delayed replay criteria: is the event important enough to be shown at a later time. If so, the system buffers the event and inserts the replay in the next available time slot (row-2 and 3 where a delayed replay for event-1 is inserted at a later time slot). Row-4 shows the generated video after replay insertion.

2.3 Audio End-Points (Tennis): Applause and Ball-Hits

In practice, the CFT approach works well for noisy soccer audio but would appear to be overkill for “cleaner” signals like tennis audio: the crowd is usually quiet during play and loud cheers and applause generally follow every point won. In fact, our experiments have shown that a good indicator of the quality of the game point, and therefore its highlight-worthiness, is in the *duration* of the applause/cheers. In our MFCC implementation, 40 filters evenly spaced over 50Hz to 3200Hz (6 octaves)

computes the mel-spectra on which the first 13 DCT coefficients are used as a vector input to a neural network. Training data is manually cropped and labeled as either Applause or Ball-hit. A window size of 100msec in steps of 50msec is used to generate MFCC vectors for the Back-propagation learning algorithm. On play sequences, ball-hit frames are differentiated from the silent frames using a simple ZCR feature. It is remarkable that fairly robust results can be obtained from this simple setup on 5 full-length games (3 different Grand-Slam tournaments, ~10hours). Relevant training data are only extracted from the first 5 minutes of each video and testing results are compiled on the remaining video sequence. High recall and precision rates of >70% has been obtained on segmenting game units. Ball-hit detection has also been used to identify *long rallies*, forming another criterion for highlight selection.

3 Content Augmentation

Advances in multimedia communications have made it possible for real-time computer-aided digital effects to be introduced into video presentations. For example, PVI's famous first-down line in American football is blended onto the field using clever chroma-keying [3], achieving a realistic implant that appears to be part of the field. However, the hardware-based method is labor-intensive and expensive. In [10], a software method is reported for inserting advertising images into selected target areas in a video broadcast. Insertion areas are manually selected and simple edge/color features are extracted for matching using geometric hashing. Video frames are buffered for smoothly aggregated insertions that do not appear abruptly. Apart from billboards, reliable segmentation of other generic landmark targets in soccer (Figure 3) is also addressed in [11]. While these methods do not need prior labeling, their dependence on domain features, eg, the soccer center ellipse, is undoubtedly a limiting factor. In the remainder of this section, we explore novel methods to overcome this limitation. Figure 4 encapsulates the essential ideas in a conceptual framework.



Fig. 3. Landmark detection in soccer video and content insertion therein

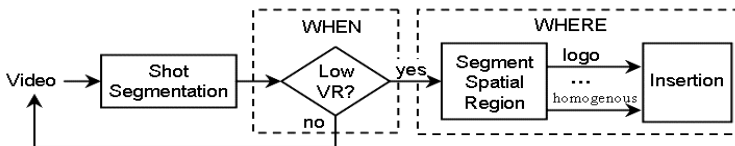


Fig. 4. A framework for Content Augmentation

3.1 Temporal Segments with Low Viewer-Relevance

The basic unit of computation in the framework remains at the shot level. Collated frames in the shot are computed to obtain a viewer-relevance (VR) measure. It is not hard to conjecture a fairly accurate VR of a specific game. For instance, in soccer, the general opinion is likely to correlate a play build-up in or towards the goal area as exciting/relevant. The equivalent to tennis is probably the game point unit, commencing from serve to out-of-bound. Basing an insertion decision upon the VR of a video shot segment facilitates the intuitively appealing notion of minimal (or none at all) disruption to viewer’s enjoyment of the game. The underlying question is to ask WHEN to insert. In this regard, the play-break classification in [5] may be usable for soccer insertion, while the applause detection methods (Section 2.3) are applicable for tennis insertion. Every image frame within each video shot that passes the criteria for low VR subsequently undergoes spatial region segmentation for content insertion.

3.2 Spatial Regions with Low Viewer-Relevance

In general, several heuristics guide the design of a spatial region segmenter for insertion. Region attributes such as homogeneity, texture similarity, motion and clutter are all factors to assign a degree of relevance to a region. For instance, the watermark indicia and time/score annotation used by many sports content distributors/broadcasters occupy small screen space, and are placed at the 2 upper corners to be of least visual disruption. One would then naturally project that these positions would also offer the “best” locations for any new content to be inserted further downstream.

On the other hand, a general characterization of a region VR appears plausible. A viable postulate is regions in a fast motion scene, eg, from a panning camera tracking a player, should have high VR. This means that even if the current shot qualifies for insertion with a low VR in the first WHEN-decision, it should arguably still fail the WHERE-decision. In contrast, regions that are uniform in terms of certain visual homogeneity metric qualitatively obtainable from the images are arguably better candidates because they contain less information and would appear to be less prominent to the human eye. The above considerations motivate the following techniques.

3.2.1 Static Region (TV Logo/Graphical Annotation) Segmentation

In general, static graphical insertions in TV are opaque or semi-transparent. The first intuition is to apply variance analysis on pixel intensity to obtain a static region mask. But this approach is sensitive to the fluctuation due to video digitization noise. Instead, we adopt a gradient-based approach by computing edge change over successive frames. A preliminary static mask S_i is obtained via time-averaging:

$$S_i = G_i + S_{i-1}, \quad (1)$$

where G_i is the gradient image of current frame i , α denotes a decay factor set to 0.7. To cope with spurious holes, eg, from the timer-ticker, morphological operations are then applied based on an elongated kernel designed to work with most TV logos and graphical annotations. Figure 5 shows examples of static masks obtained.

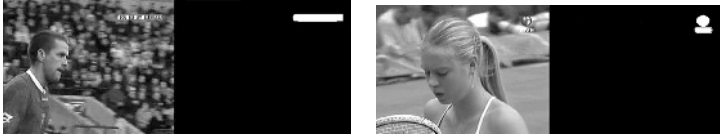


Fig. 5. Static mask examples on soccer and tennis

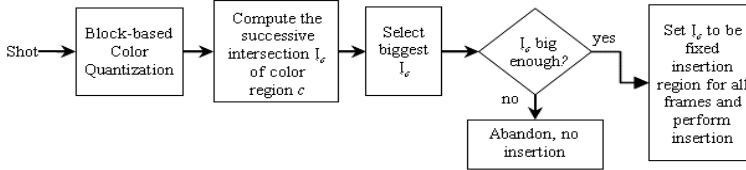


Fig. 6. Homogeneous region segmentation

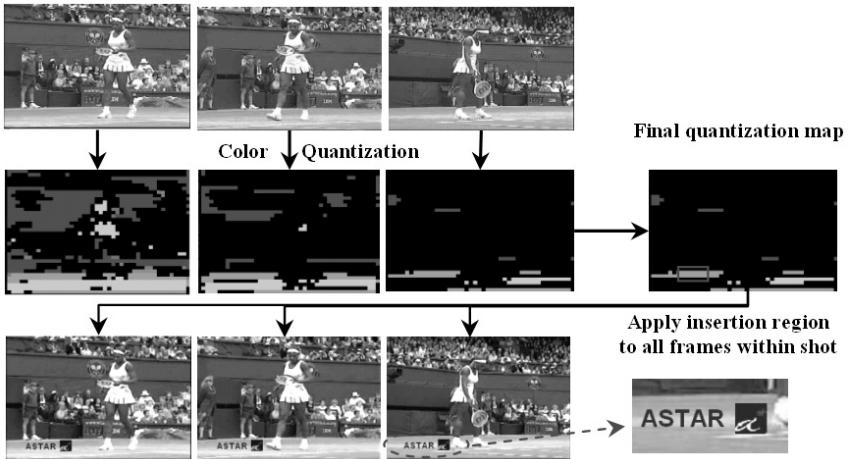


Fig. 7. Top-row: 3 original images successively spaced 1-sec apart. Row-2: evolution of the color quantization map; Row-3: Insertions (logo) onto all frames within shot

3.2.2 Homogeneous Region Segmentation

Figure 6 shows an exemplary flow-chart for segmenting a color-homogeneous region for insertion. Image frames in the shot are first divided into 32×32 non-overlapping blocks. These are then collated to be quantized [12] into a small number of colors (eg, 8). The quantized color indices are then used to isolate regions with color consistency. This is done by computing the boundary extent of each color-coded region using simple connected component. As the frame content changes, the boundary extent of these regions will also change. By taking the intersection of all regions computed within the shot, we derive a rectangular region which has maintained its color integrity for the duration of the shot. A decision is then made as to whether the region is usable for insertion. This may be based on the match of its dimension to the geometry of the content to be inserted: a small square region is useful for a logo insertion (Figure 7), while an elongated one can be used for an animation sequence (Figure 8).

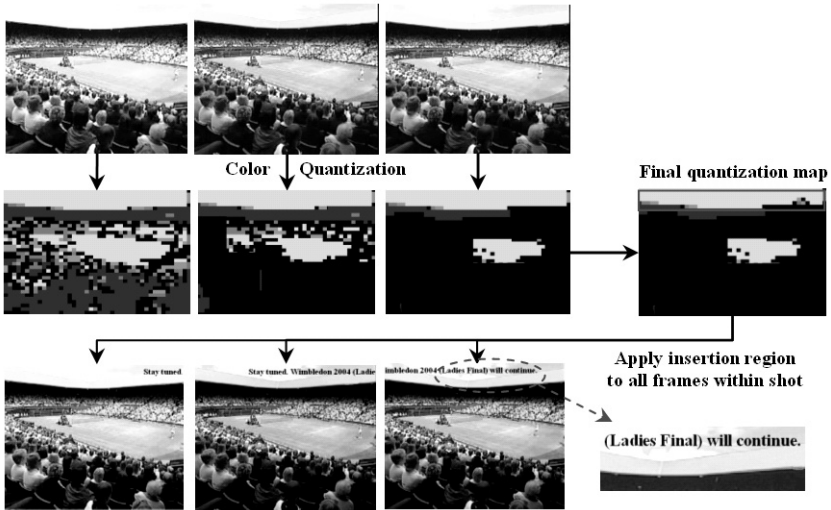


Fig. 8. Text insertion on elongated regions in original images 2-sec apart

4 Experimental Results

To verify the validity of our approach, we conduct subjective viewing tests using highlights that are automatically generated from a soccer video and a tennis video. Content augmentation described in Section 3 is then applied to these highlights. Since not every highlight segment has a successful insertion, we arbitrarily chose 9 soccer highlights of which 8 has insertions, and 10 tennis highlights of which 8 has insertions. This makes a total of 16 insertions on the 19 highlights. Only static logo insertions are used, and these are uniquely taken from famous trade-marks such as Nike, Mastercard, etc. To throw the subjects off guard, they are first asked to identify some trivia in the video before viewing starts. After the viewing, they are then asked to recall any logos they have noticed (subtly), and whether these exposures have reduced their viewing experience (acceptability). The cumulative results over 8 viewers are tabulated in Table 1. Most gave a high opinion of the quality of our highlights, and did not notice the pattern of advertising insertions only after quite a few of them have appeared. We also have a general concurrence on our design of making advertising insertions occur during play-break, in order to minimize interference with the game proper. The minority opinion of opposition came in 3 forms: (1). No insertions at all; (2). Restrict the insertions to a single place; (3) Legal concern on insertions over the static watermark logo.

Table 1. Subtlety and Acceptability of Insertions

Sports	Subtlety	Acceptability
Soccer	50%	50%
Tennis	40%	60%

On the whole, Column 2 and 3 show a high level of acceptance and high recall of the exposure. This must be good news for advertisers.

5 Conclusion and Future Work

We expect that automatic sports highlight technology will facilitate alternative channels for sports content distribution and broadcasting. A framework for augmenting video highlights with attendant content is developed in this paper. The obvious use of this is in product branding/advertising, a traditional financial pillar for broadcast media. Based on the viewing patterns, it is also easy to generalize the framework to incorporate means for mining purchasing preferences for target advertising.

References

1. Adami, N., Leonardi, R., Migliorati, P.: An Overview of Multi-modal Techniques for the Characterization of Sport Programmes. In: Proc. SPIE – VCIP 2003, pp. 1296-1306
2. Wang, J., Xu, C., Chng, E., Wan, K., Tian, Q.: Automatic Highlight Detection and Replay Generation for Soccer Video. Full paper to appear in ACM Multimedia 2004.
3. PVI Virtual Media Services: <http://www.pvimage.com/pvi/index.html>
4. Duan, L., Xu, M., Chua, T., Tian, Q., Xu, C.: A mid-level representation framework for semantic sports video analysis. In: Proc ACM Multimedia 2003, pp. 33-44
5. Xie, L., Chang, S., Divakaran, A., Sun, H.: Structure Analysis of Soccer Video with Hidden Markov Models. In: Proc ICASSP 2002, Orlando, FL, USA
6. Sudhir, G., Lee, J., Jain, A.: Automatic classification of tennis video for high-level content-based retrieval. In: Proc CAIVD, 1998, pp. 81 -90
7. Xiong, Z., et.al, T.: Audio events detection based highlight extraction from baseball, golf and soccer games in a unified framework”, In: Proc of ICASSP 2003, Vol V pp 632-635
8. Wan, K., Xu, C.: Robust Soccer Highlight Generation with a Novel Dominant-Speech Feature Extractor. In: Proc ICME 2004, Taiwan.
9. Wan, K., Xu, C.: Efficient Multimodal Features for Automatic Soccer Highlight Generation. To appear in: Proc ICPR 2004, Cambridge, UK.
10. Medioni, G., Guy, et.al: Real-time billboard substitution in a videostream. In: Proc 10th Tyrrhenian International Workshop on Digital Communications, Italy, 1998, pp.71-84
11. Wan, K., Yan, X., Yu, X., Xu, C.: Real-time goal-mouth detection in MPEG soccer video. In Proc ACM Multimedia 2003, pp. 311-314
12. Gervautz, M., Purgathofer, W.: A Simple Method for Color Quantization: Octree Quantization. In: Proc. CGI '88, pp. 219-231

Audio-Assisted Video Browsing for DVD Recorders

Ajay Divakaran¹, Isao Otsuka², Regunathan Radhakrishnan¹,
Kazuhiko Nakane², and Masaharu Ogawa²

¹ Mitsubishi Electric Research Laboratories, Cambridge, USA,
ajayd@mer1.com, regu@mer1.com,

² Mitsubishi Electric Corporation, Kyoto, Japan

Abstract. We present an audio-assisted video browsing system for a Hard Disk Drive (HDD) enhanced DVD recorder. We focus on our sports highlights extraction based on audio classification. We have systematically established that sports highlights are indicated by the presence of audience reaction such as cheering, applause and the commentator's excited speech. That enables us to develop a common highlights extraction technique, based on detection of audience reaction, for five different sports, viz. soccer, baseball, golf, sumo wrestling and horseracing. Our extraction accuracy is high. Furthermore, the percentage duration of the audience reaction gives us a simple importance measure for each of the highlights. We can then get a summary of any desired length by appropriately choosing a threshold for the importance measure. We process the AC-3 audio directly thereby enabling simple integration of our technique into our target platform.

1 Introduction and Motivation

Our target application is the Personal Video Recorder (PVR) (See Nakane et al [2003]). Current PVR's can store up to 200 hours of video content, and the storage capacity is expected to further grow in the future. There is therefore a great need for video browsing techniques that help the user skim through many hours of content quickly, as well as to select which part of the content to play in full. While video browsing and summarization (See for example Divakaran et al [2003]) has been an active area of research, realization of video summarization techniques on consumer video devices is an open challenge. First, there is no constraint on the variety of content genres that the consumer will record, so the summarization algorithms have to be applicable to a wide variety of content. Second, consumer video devices typically use partly-programmable custom integrated circuits that offer limited computational resources.

In this paper, we propose a common framework based on processing the audio to gauge audience reaction for extraction of highlights of five sports, Soccer, Baseball, Golf, Sumo wrestling and horse racing, which is also applicable to other sports in which audience reaction is indicative of an interesting event. Since the compressed audio coefficients are in the frequency domain, our audio

classification directly works on them thus obviating the computation required for conversion to the frequency domain. We thus take advantage of the target platform architecture to vastly reduce the computational complexity. The rest of the paper is organized as follows. In section 2, we describe the audio classification framework. In section 3, we describe the highlights extraction based on detection of audience reaction and the ranking of highlights and present some experimental results. In section 4, we describe the implementation of the proposed algorithm on the target platform. In section 5, we present our conclusions and possibilities for further research.

2 Audio Classification Framework

The following sound classes span almost all of the sounds in sports domain: Applause, “The commentator’s Excited Speech combined with Cheering,” “Cheering”, Music, Speech & Speech with Music (See Xiong et al [2003]) . We collected training data from several hours of broadcast video data for each of these sound classes. We extract MDCT (Modified Discrete Cosine Transform) coefficients from the AC-3 encoded audio, at the rate of 32 frames per second, with each frame containing 256 coefficients. The 256 coefficient frame is considered to be our feature vector and is normalized using the audio energy in the last second of the clip. We reduce the dimension of the 256 dimensional feature vector through Principal Component Analysis (PCA), to 22. We trained Gaussian Mixture Models (GMMs) to model the distribution of features for each of the sound classes. The number of mixture components were found using the minimum description length principle so as to get high classification accuracy (See Xiong et al [2004]). Then, given a test clip, we extract the features for every frame and assign a class label corresponding to the sound class model for which the likelihood of the observed features is maximum. We illustrate our audio classification in Figure 1.

Our audio classification techniques are robust in the face of the extremely noisy ambience of broadcast sports video. Such robustness stems from the careful collection of the training data as well as the use of the Minimum Description Length principle while training the GMM’s. However, the accuracy is clearly not perfect, and hence we have to compensate for the mis-classifications when we extract the highlights.

3 Sports Highlights Extraction

Our sports highlights extraction is guided by the intuition that interesting events lead to notable audience reaction in the form of applause or cheering combined with excited speech from the commentator. Furthermore, the more interesting the event, the longer the audience reaction to the event. For example, a goal scoring move in soccer would get an audience to cheer and the commentator to comment excitedly for a much longer time than would a mildly interesting moment such as a free kick. Therefore, the audience reaction provides a powerful indicator of interesting events or highlights across a wide range of sports. We

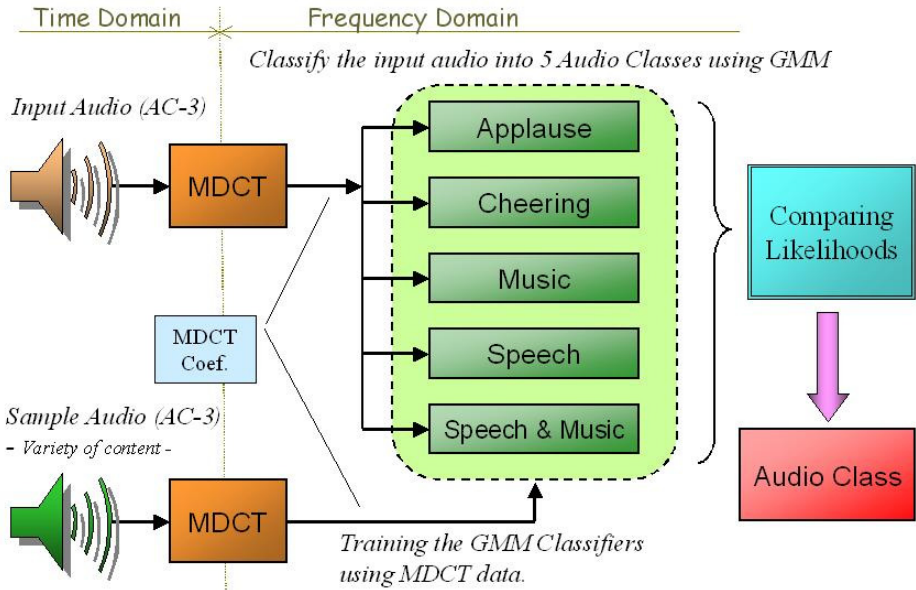


Fig. 1. Audio Classification Framework

thus have a genre-independent way to extract sports highlights as illustrated in Figure 2. We compute the percentage of the key audio class such as applause or excited speech and cheering in a sliding window of a certain pre-set length such as 10 seconds. Note that unlike our previous work (Xiong et al [2003]) we do not use the length of contiguous segments of unbroken audience reaction. We use the percentage instead to compensate for mis-classification that falsely breaks up long stretches of audience reaction. Furthermore, we weight the percentage of audience reaction in the window with the audio energy. We do so to eliminate false alarms caused by low energy audio that is mis-classified.

As shown in Figure 3, our sports highlights extraction gives rise to an importance measure for every time unit of the video sequence. We can then get a summary of a desired length by setting the threshold for the importance appropriately. Thus for example, if we want just the few most important highlights, we would increase the threshold until the right number of highlights are above the threshold.

We have tried our approach with a wide variety of sports content drawn from Soccer, Baseball, Golf, Sumo Wrestling and Horse-racing. We have got satisfactory accuracy with 30 odd games from Japanese and U.S. broadcast sports content. We illustrate a typical result with a Golf game in Figure 4. We determine the accuracy by first manually annotating notable events such as goals, home runs etc., and then checking to see if our automatic technique detects them. We find that we have very few misses when we aim for setting the summary duration to roughly ten percent of the total duration of the content.

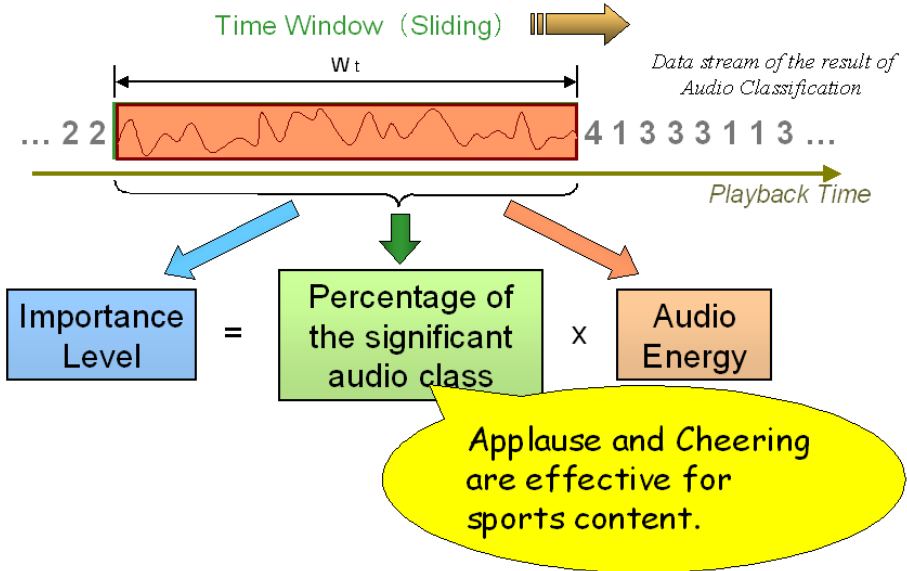


Fig. 2. Sports Highlights Extraction Framework

However, note that since the threshold is variable, the user can adjust it to get better results. We believe the reason for the good performance is that audience reaction is a reliable indicator of how notable the event is. The audience reaction is in fact an indication of the consensus among the human observers of the event on how remarkable the event is. Therefore, if the audio classification works well, the highlights extraction is very successful. We also find that with sports such as golf that have low ambient sound because the spectators are relatively quiet, we get much better results than with sports such as soccer in which the audience is noisy throughout. However, the excited speech of the commentator is spectrally so distinctive, that it is detectable even when the ambience is noisy.

4 Realization on Target Platform

Since our sports highlights extraction uses the same essential algorithm for a wide variety of sports content, it requires only a simple enhancement of the target platform viz. the DVD recorder. We illustrate our proposed enhancement in Figure 5. Furthermore, our use of the Minimum Description Length principle while training GMM's gives rise to compact GMM's with as few mixture components as is feasible while maintaining accuracy. Notice also that our direct use of the AC-3 coefficients makes our enhancement a simple and direct modification of the audio encoder.

The summarization meta-data is generated during the recording phase as can be seen in Figure 5. Since we use a sliding window, there is a delay equal to the

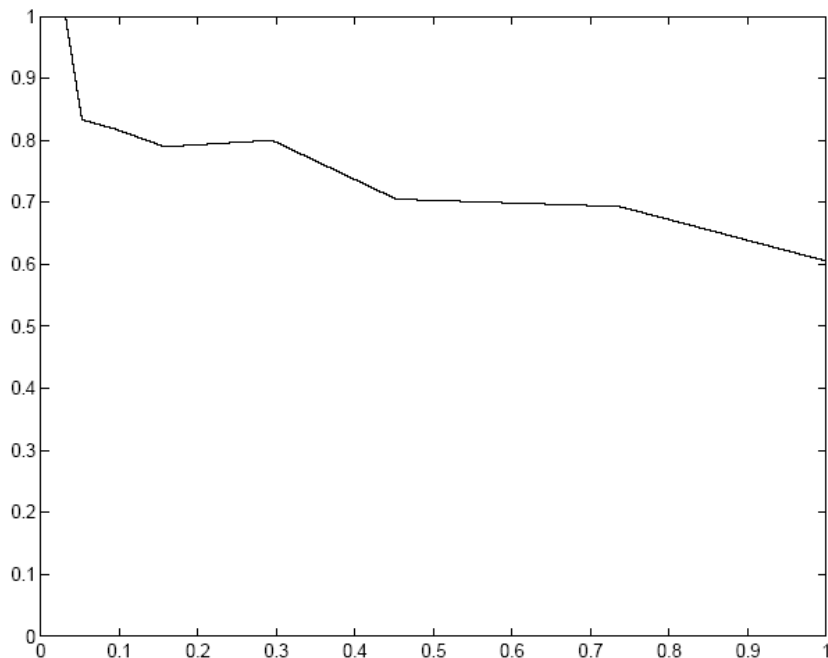


Fig. 3. Precision (y) -Recall (x) Plot for British Open Golf Game

length of the window after the recording is completed. However, such a delay is only a few seconds, hence the summarization meta-data is available to the user almost as soon as the recording is complete. The meta-data is written out to the HDD or to the DVD disc depending on the user preference. Our meta-data syntax is simple since it is merely an importance measure profile of the video as illustrated in Figure 3. It therefore takes up very little space and can be written in the private space of the DVD disc.

5 Conclusion

We presented a common highlights extraction algorithm for a wide variety of sports using audio detection of audience reaction to notable events, that has reasonable accuracy. We proposed an approach to realization of the algorithm on our target platform. Note that our algorithm design from the outset ensured simple integration into the target platform. We will present a demonstration of our algorithm at the conference.

There are several avenues for further improvement of the system. First, we will explore improvement of the audio classification accuracy. Second, we will

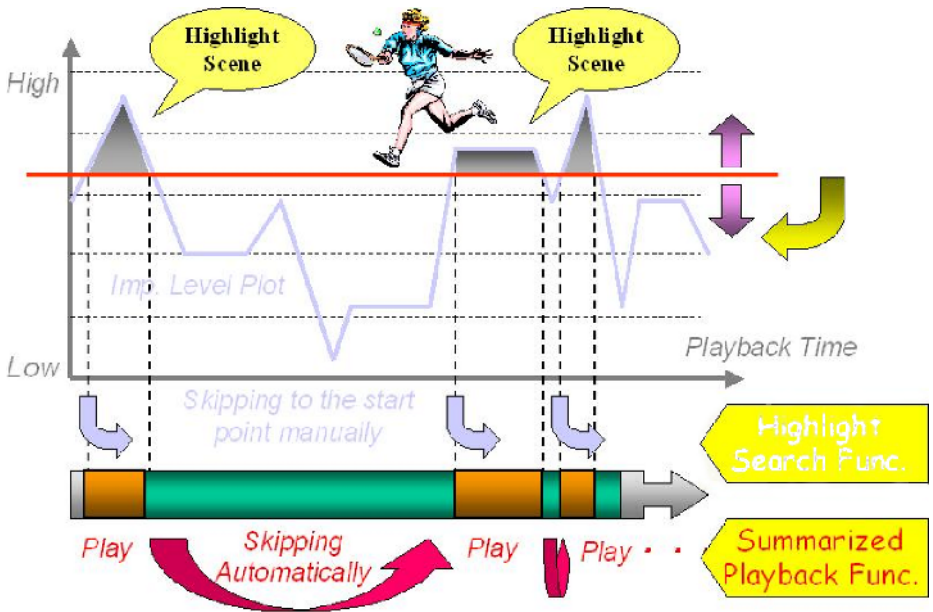


Fig. 4. Scalable Summarization

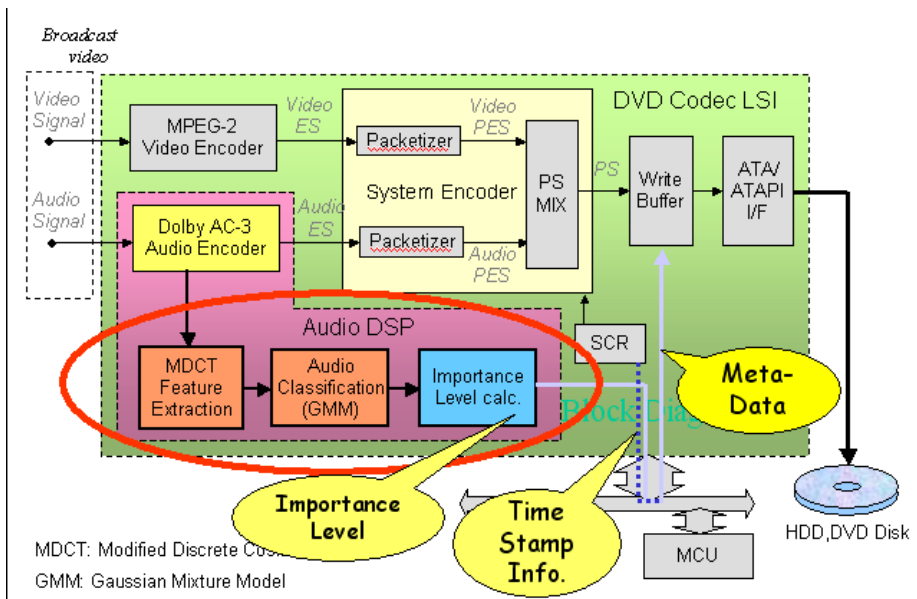


Fig. 5. Realization on Target Platform

explore incorporation of visual cues into the system. We hope to eliminate false alarms as well as carry out a richer segmentation of the content. Third, we will extend our framework beyond sports video. Our emphasis has been on “infotainment” content such as sports and news because we believe that summarization is best suited for browsing for information. However, we now need to find out how to summarize other genres such as dramas and documentary movies, and how to accommodate our extension of the algorithms on our target platform. Fourth, we plan to explore the user interface. In our view, the user interface is the critical component of the system that can much more than compensate for lack of accuracy in detecting sports highlights. We need to develop a user interface that will work well in our target platform which is remote-control driven.

References

- [2003] A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong and R. Cabasson, *Video Summarization using MPEG-7 Motion Activity and Audio Descriptors*, in Video Mining, A. Rosenfeld, D. Doermann, and D. DeMenthon, Eds., Kluwer Academic Publishers, 2003.
- [2003] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, *Audio Events Detection based Highlights Extraction from Baseball, Golf and Soccer Games in A Unified Framework*, ICASSP 2003, April 6-10, 2003.
- [2003] K. Nakane, I. Otsuka, K. Esumi, T. Murakami and A. Divakaran, *A Content-based Browsing System for HDD and/or recordable DVD Personal Video Recorder*, IEEE Conference on Consumer Electronics (ICCE), 2003.
- [2004] I. Otsuka, A. Divakaran, K. Nakane, and M. Ogawa, *HDD Enabled DVD Recorder System*, IPSJ Conference, Japan, March 2004.
- [2004] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, *Effective and Efficient Sport Highlights Extraction using the Minimum Description Length Criterion in Selecting GMM Structures*, ICME 2004, June 27-30, 2004.

A Robust Image Watermarking Technique for JPEG Images Using QuadTrees

Kil-Sang Yoo, Mi-Ae Kim, and Won-Hyung Lee

Chung-Ang University,
Department of Image Engineering,
Graduate School of Advanced Imaging Science and Multimedia and Film,
221 Hukseok-Dong, Dongjak-Gu, Seoul, Korea
lucky@ms.cau.ac.kr, kimma@dreamwiz.com, whlee@cau.ac.kr

Abstract. In this paper, we propose a robust watermarking technique that embeds a Gaussian sequence watermarks into low frequency area of the wavelet transform domain. We overcome between the image degradation problem and the watermark robustness by embedding watermarks into visually insensitive pixels using QuadTrees. We determined embedding pixels according to the picked 1 by 1 block in a sparse metrics which is decomposition DWT low band. Decomposed Sparse matrix areas must not degrade after JPEG compression. We compare our experimental results with respect to JPEG compression with Cox's and Joo's popular correlation-based method. The experimental results show that the proposed algorithm successfully survives several kinds of image processing operations especially for JPEG lossy compression.

Keywords: Watermarking, QuadTrees, DWT, Watermark, Robust Watermarking.

1 Introduction

Due to the rapid and extensive growth of multimedia network system, data can be distributed much faster and easier than before. The protection and enforcement of intellectual property rights for digital media have become an important issue, and many watermarking techniques have been developed. The Digital image watermarking technique is embedding watermark into digital image without degradation of original image. Some examples of watermark information include a logo, a picture, a signature, or sequence.

A significant number of watermarking methods have been recently reported. Most of these methods embed the watermark into the spectral coefficients of images by using signal transformation such as discrete cosine transformation(DCT) or discrete wavelet transformation(DWT) because embedding in the frequency domain is more tolerant to attacks and image processing than embedding in the spatial domain. Wavelets are becoming a key technique in the ongoing source compression standard JPEG-2000. In several recent publications, this technique has been applied to image watermarking. The positive arguments closely resemble those for advocating DCT for JPEG (i.e. preventing watermark removal by

JPEG-2000 lossy compression, reusing previous studies on source coding regarding the visibility of image degradations, and offering the possibility of embedding in the compressed domain). Our technique is to embed Gaussian sequence watermark into an image in the wavelet domain, combined with the position sequences and original image, and creates a QuadTree [1]. The QuadTree can help the processing of watermark embedding and watermark verification successfully. This paper is organized as follows. The watermark embedding and extracting approach are described in Section 2. In Section 3, the experimental results are shown. The conclusions of our study are stated in Section 4.

2 The Proposed Watermarking Technique

A basic block diagram of a watermarking system is illustrated in Fig. 1. The original image can represent some transform domain signal such as the DWT coefficients. The watermark sequence which is a Gaussian sequence of owner number. Then watermark is embedded according to PRNG (pseudo random number generator). That is to say, we created PRNG using seed key. And the watermark sequence permuted according to the created PRNG. We determined embedding pixels according to the picked 1 by 1 block in a sparse metrics which is decomposition DWT low band.

In the watermark detection procedure, we perceive the watermark distribution location in the reverse method. Namely, we recreate PRNG by applying seed key used in the embedding procedure. We will describe the watermarking procedure in detail as the following.

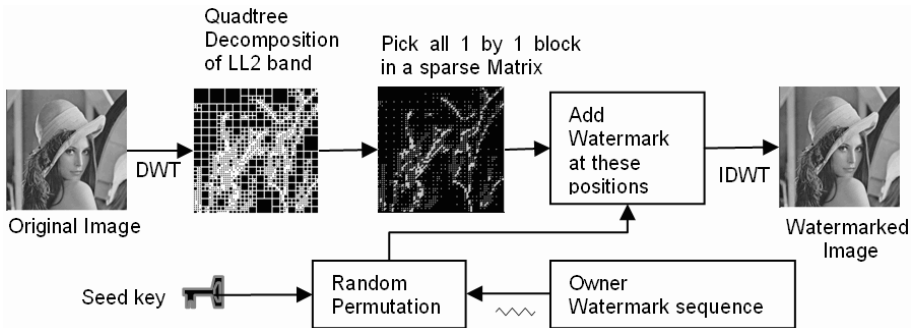


Fig. 1. Block diagram of a watermark embedding procedure

2.1 Watermark Embedding

The wavelet transform is identical to a hierarchical sub band system, where the sub bands are logarithmically spaced in frequency. An image is first decomposed

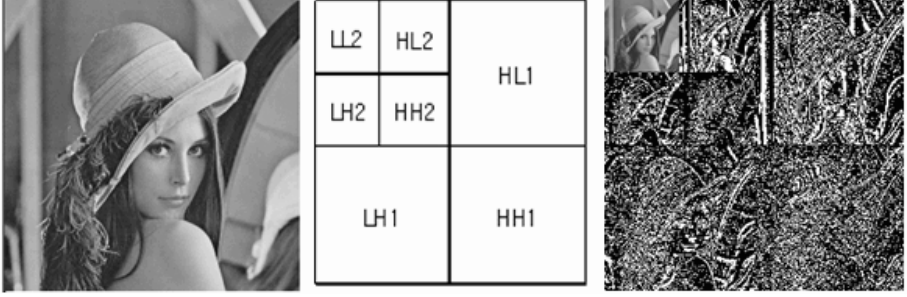


Fig. 2. DWT two-level wavelet decomposition of an image

into four parts of high, middle, and low frequencies (i.e., LL1, HL1, LH1, HH1 sub-bands) by critically sub-sampling horizontal and vertical channels using sub-band filters as (2). To obtain the next coarser scaled wavelet coefficients, the sub-band LL1 is further decomposed and critically sub-sampled. We used the Haar wavelet transform that is a kind of discrete wavelet transform [2]. An original gray Lena image and its DWT decomposition are shown in Fig. 2.

The watermark is embedded into low frequency band (LL2) of the two-level wavelet decomposition. The determined LLn can be seen as a reduced version of the original image. In our embedding strategy, an LLn is decomposed by QuadTree to find pixels which is watermark embedding position.

$$(W_{\Psi})(a, b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} f(t) \Psi \left(\frac{t-b}{a} \right) dt \quad (1)$$

QuadTree is a well known data structure, which can be used to describe the spatial information of an image. It is also used variety of image analysis and compression applications. QuadTree decomposition is subdividing an image into blocks that contain "similar" pixels. It works by successive refinement of blocks. For example, suppose the input image is 128 by 128. It starts with a single 128 by 128 block. If the pixels in the block are not similar, it subdivides the block into four 64 by 64 blocks, and then it subdivides the non-similar 64 by 64 block into four 32 by 32 blocks, and so on. After decomposed by QuadTree, it returns the QuadTree decomposition as a sparse matrix which is shown in the (c) of the Fig. 3. This representation lends itself to determine the watermark embedding locations of all 1-by-1 blocks in the QuadTree decomposition. Sparse matrix areas must not degradation after JPEG compression as Fig. 3 (d).

We use a Gaussian sequence watermark w , where w sequence is chosen according to independent normal distributions with standard deviation σ . A secret key is chosen as the seed of predefined pseudo random number generator (or sometimes called as random permutation) [4]. The Gaussian sequence watermark is shuffled by using PRNG. The LL band is modulated according to the formula (2) until watermark length times.

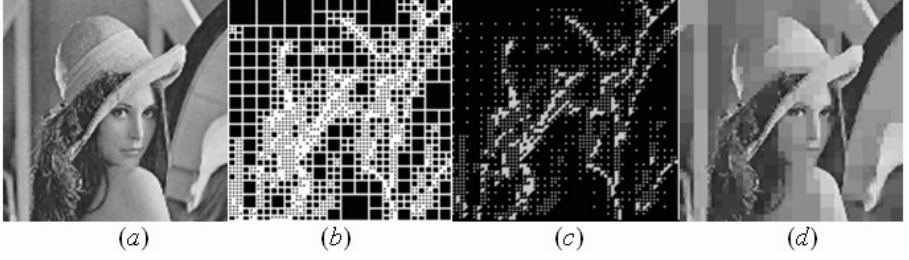


Fig. 3. Decomposition process of QuadTree; (a) Original Lena image, (b) Subdivides a Lena image of LL2, (c) returns the QuadTree decomposition as a sparse matrix, (d) After decomposed image

$$LLn(positionV(i)) = LLn(positionV(i)) * (1 + sf * w(prng(i))) \quad (2)$$

The $positionV$ denotes the embedding pixel of a sparse matrix. Where w is the watermark bits, and sf is a scaling factor which is the watermark strength and can be adjusted to achieve a reasonable compromise between the robustness of the watermark and its visibility. Large values of sf lead to more robust schemes but the watermark becomes more visible because DWT coefficients are modified by a larger amount. Finally, the watermarked image is obtained by applying the inverse DWT (IDWT) to the coefficients.

2.2 Watermark Detection

We proposed techniques that need the original image for the watermark extraction are called non-blind watermarking. Typically, such schemes are more robust than blind schemes that do not need the original image for watermark extraction. The same wavelet decomposition used in the embedding is applied to both the original and embedded images. The watermark embedding locations are obtained by QuadTree decomposition from original image. Watermark detection is done by subtracting the original image from a suspected image, calculating the DWT of the difference, and extracting the watermark sequence according to the position sequence generated by PRNG with security "key" which is used in embedding procedure. These positions information can be referred to as a "key" with much the same meaning as is used in cryptography.

The watermarked image quality is measured using PSNR (Peak Signal to Noise Ratio) by the equation (3) where MSE is the mean-square error between a watermarked image and original image [4]. In this comparison, we use the similarity measure given in (4).

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} = 10 \log_{10} \frac{255^2}{\sum (f(x) - f(x'))^2} \quad (3)$$

$$Sim(w, w') = \frac{w \cdot w'}{\sqrt{w' \cdot w'}} \quad (4)$$

3 Experimental Results

We demonstrated the robustness of the proposed watermarking algorithm by using the MatLab 6.1. In order to evaluate the proposed watermarking scheme, we took the gray Lena image of 512 by 512 pixels. The Gaussian sequence watermark is generated from the seed number 250 and its length of 1000 used. We added the watermark to the image by modifying 1000 of the more perceptually significant components of the image using QuadTree decomposition. A two-level discrete wavelet transform is employed and thus the size of LL band to be embedded is 128 by 128. More specifically, the 1000 largest coefficients of the DWT were used. A fixed scale factor of 0.08 was used throughout.

Our method is compared with Cox's method [5] and Joo's method [6]. The resulting PSNRs were 39.76 dB for Cox's, 40.69 dB for Joo's and 40.44 dB for ours. For both methods, the original watermark was compared with the one extracted from the JPEG-compressed images. Fig. 4 shows the detection responses between the watermark sequence extracted from the JPEG-compressed image (10% quality factor) and the watermark sequences set generated from seed numbers taken by 1 to 1000. Each peak value in the figure is matched with the similarity value corresponding to the same quality factor shown in Fig. 5. As shown in the figures, the proposed method suffers much less from JPEG lossy compression than Cox's and Joo's method.

The proposed method was also tested with respect to non-geometric attacks provided from CheckMark 1.2 [7]. The experimental results show clearly detected against dithering, JPEG compression, filtering, remodulation and JPEG2000 compression without copy attack and thresholding attack as shown Table 1. As

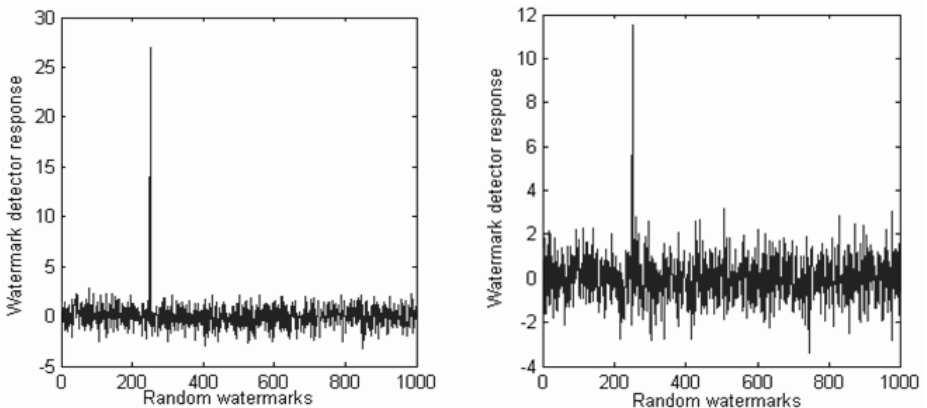


Fig. 4. Watermark detector response to 1000 randomly generated watermarks. Only one watermark (the one to which the detector was set to respond) matches; Left: Detector response to JPEG compression quality 10% and Right: Detector response to JPEG compression quality 1%

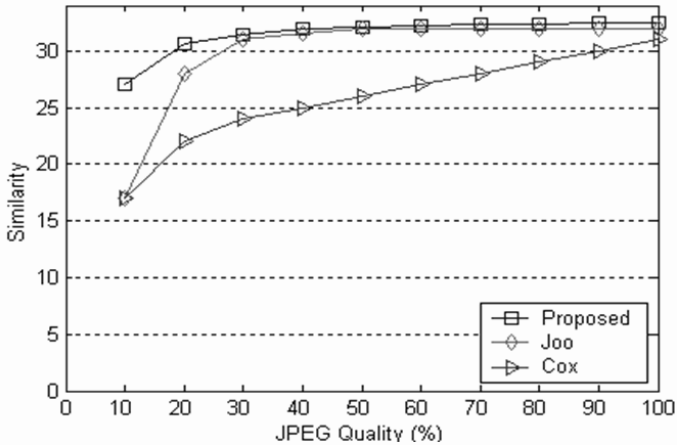


Fig. 5. On the Robustness Compared to Other Approaches; The proposed embedding scheme is more robust against than other approaches

Table 1. Detection results using StirMark 1.2x

Attack(s)	im1	im2	im3	im4	im5	Total
MAP(6)	6	6	6	6	6	30
JPEG compression(12)	12	12	12	12	12	60
Filting(3)	3	3	3	3	3	15
Wavelet compression(10)	10	10	10	10	10	50
ML(7)	7	7	7	7	7	35
Remodulation(4)	4	4	4	4	4	20
reSample(2)	2	2	2	2	2	10
Copy attack(1)	0	0	0	0	0	0
ColorReduce(2)	0	1	1	1	0	3
Total(47)	44	45	45	45	44	223

Table 2. Detection comparisons to other techniques using StirMark 1.2

Methods	Non-geometric (5 images, 235 attacks)
Wang [8]	74%
Cox	90%
Xia [9]	84%
Kim [10]	48%
Joo	93%
Proposed	95%

you can see, proposed method survivals 223 among 235 attacks. Table 2 shows total detection ratio comparisons to other techniques.

4 Conclusions

Most watermarking techniques embed watermarks in the middle frequency range for robustness with invisibility. The low frequency range is not suitable for embedding because of severe visual degradation. In this paper, we present the algorithm that reduces the degradation of a watermarked image by embedding the watermark sequence into the low frequency bands of the wavelet transform domain. The watermark sequence was permuted according to the created PRNG. We determined embedding pixels according to the picked 1 by 1 block in a sparse metrics which is decomposition DWT low band, it picks the strong coefficients to embed watermarks. These coefficients are the most important and strong parts of the whole image, so it is not easy to be changed after the JPEG or JPEG-2000 lossy compression.

The experimental results show good robustness against JPEG compression because the sparse matrix areas must not degrade after JPEG compression as seen. We showed that our algorithm can extract a reliable copy of the watermark from imagery degraded with signal processing procedures such as filtering and noise addition. Besides, in the robustness compared, we show that the proposed embedding scheme is more robust against non-geometric attacks than other approaches.

In Future work, we will extend the proposed scheme to blind watermarking algorithm designed to lost watermark reconstruction.

Acknowledgment. This research was supported by the Ministry of Education, Seoul, Korea, under the BK21 project and research fund of Chung-Ang University in Seoul.

References

1. Rafael C. Gonzalez, Richard E. Woods.: Digital Image Processing. Addison-Wesley Publishing Company, (1993) 461–464
2. Albert Bogges Francis J. Narcowich.: A First Course in Wavelets with Fourier Analysis. Prentice Hall, Upper Saddle River, (2001) 155–182
3. Jui-Cheng Yen: Watermark embedded in permuted domain. Electronics Letters, Vol. 37, Issue: 2, 18 Jan (2001) 80–81
4. Stefan Katzenbeisser, Fabien A. P, Petticolis.: Information hiding techniques for steganography and digital watermarking. Artech House, Boston London (2000) 128–129
5. I.J. Cox, J. Killian, T. Leighton, and T. Shamoan.: Secure Spread Spectrum for Multimedia. IEEE Trans. on Image Processing, vol. 6, no. 12, (1997) 1673–1687
6. Sanghyun Joo , Youngho Suh , Jaeho Shin , Hisakazu Kikuchi and Sung-Joon Cho.: A New Robust Watermark Embedding into Wavelet DC Components. ETRI Journal, vol. 24, no.5, Oct. (2002) 401–404

7. CheckMark, <http://watermarking.unige.ch/Checkmark/>
8. Houngh-Jyh Wang, Po-Chyi Su, and C.-C. Jay Kuo.: Wavelet-based digital image watermarking. *Optics Express*, December (1998) 497
9. X. Xia, C.G. Bonchelet, and G.R. Arce.: A Multiresolution Watermark for Digital Images. *Proc. of IEEE ICIP*, Santa Barbara, CA, USA, Oct. (1997) 548–551
10. J.R. Kim and Y.S. Moon.: A Robust Wavelet-Based Digital Watermarking Using Level-Adaptive Thresholding. *Proc. of IEEE ICIP*, vol. 2, Kobe, Japan, Oct. (1999) 226–230

A Fragile Watermarking Technique for Image Authentication Using Singular Value Decomposition

Vivi Oktavia and Won-Hyung Lee

Department of Image Engineering
Graduate School of Advanced Imaging Science, Multimedia & Film
Chung-Ang University
vivi_o@hotmail.com, whlee@cau.ac.kr

Abstract. In this paper, we propose a block SVD-based fragile watermarking technique for image authentication. Using this technique, we can detect any modification made to the image and indicate the specific locations where the modification was made. We utilize the singular value of image block as the authentication data. The image is divided into same size blocks, and authentication data is inserted into the LSB plane of each block. Authentication data is produced from XOR operation between watermark image and binary bits obtained from singular value of one block image. Singular value is converted to binary bits using modular arithmetic. Security of this technique resides on two keys, so in the case where the keys are incorrect, the extraction process will return an image that resembles noise, the same case as if the image is not watermarked, or the watermarked image is modified. If the unmodified watermarked image is used, the extraction process will return the correct watermark image.

1 Introduction

An old proverb said pictures don't lie, but pictures do lie after some image editing software was released. Now, problems arise in questioning whether an image is really authentic and credible to be used as evidence, for example, in the court, in insurance company, in newspaper and television, etc. Manipulation can easily be done with sophisticated image editing software, so it is generally impossible to judge whether an image is authentic only by looking at it. To address this issue, fragile watermarking has been proposed for authentication purposes.

Watermarking is a technique to embed additional data or signal (called watermark) into multimedia data. Fragile watermarking has the same principle, embedding watermark to a digital data, except that the watermark is likely to become undetectable if the data is modified in any way. An example of earliest fragile watermarking techniques is embedding watermark to the LSB (least significant bit) of the image that does not cause visual artifacts to the image. But such LSB manipulation is not secure against malicious attacks, because it is possible to alter an image without changing the LSB value. Yeung-Mintzer [1]

proposed an authentication method using key-dependent binary-valued function (lookup-table) to modify the pixels value in order to embed a watermark, and deployed error diffusion method to maintain proper average color. Although this technique can detect any changes in pixel values, it does not resist to impersonation attack if the same lookup table and watermark are used for multiple watermarked images [2]. Wong-Memon [3] proposed a secret and public-key image watermarking scheme for image authentication. They used cryptographic hash function such as MD5, which makes the implementation of this scheme slow. Byun et. al. [4] used singular values (SVs) of an image matrix as authentication data. To extend the algorithm for localization, the SVs and authentication logo are tiled and inserted to LSB of the original image, so it will be able to localize tampered blocks. Unfortunately, manipulation of pixels value may change all the SVs, and tiling operation will further propagate the SV error to all block watermarks. As the result, it is not always possible to localize the tampered regions.

In this paper, we propose a block SVD-based fragile watermarking technique. The authentication data is produced from XOR operation between the watermark image and binary bits obtained from the singular value of the image. Using modular arithmetic as in [4], binary bits are generated from the singular value of one image block. The localization property is obtained by dividing the original image into same size $s \times s$ blocks, and embedding the authentication data to LSB plane for each block. The security of this technique resides on two keys used in the embedding process. The first key, which consists of s^2 bits, is used to replace the LSB of each image block before calculating the SVs value. After the SVs are obtained, they are tiled in order to get the same size authentication data as the image block, and then random permutation is applied. The second key is used as a seed for random permutation. If wrong key is used in the extraction process, it will return image, which resembles noise, not the watermark image.

2 Singular Value Decomposition

Any $m \times n$ real-valued matrix A , with $m \geq n$ can be written as the product of three matrices

$$A = USV^T \quad (1)$$

The columns of the $m \times m$ matrix U are mutually orthogonal unit vectors, as are the columns of the $n \times n$ matrix V . The $m \times n$ matrix S is a pseudo-diagonal matrix, where its diagonal elements are

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0 \quad (2)$$

and it is called the singular value (SV) of A . While both U and V are not unique, the singular values σ_i are fully determined by A . From the viewpoint of image processing application, singular values (SVs) represent intrinsic algebraic image properties [5].

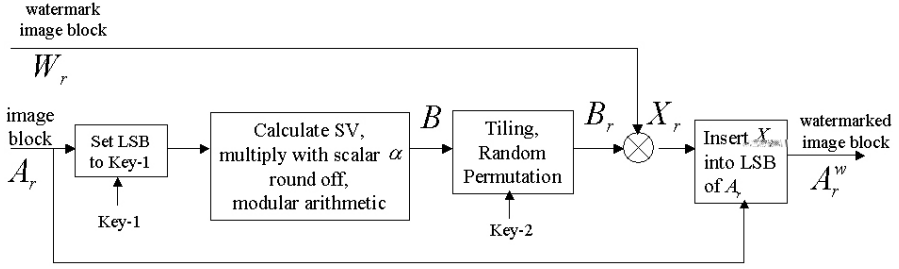


Fig. 1. Block diagram of watermark embedding process

3 Watermark Embedding

Consider a grayscale image A of size $n \times n$ pixels as the original image, and a binary image W , which has the same size as A , which will be used as the watermark. Note that a digital image can be considered as a non-negative matrix. The embedding technique is detailed as follows:

First, we divide the original image A and the watermark image W into $s \times s$ square blocks. Let us assume the original image size to be 512×512 pixels, and it is divided into 4096 image blocks where each image block size is 8×8 pixels. Denote A_r to be the image block and W_r as the watermark block where $r = 1, 2, \dots, 4096$. Note that r is the index block. Then, a 64-bits Key-1 is generated, which has the same size as the numbers of pixels in each image block A_r .

For each block A_r , do the following process:

1. Replace the LSB of A_r with Key-1.
2. Calculate singular values (SVs) of A_r . Because A_r is 8×8 size, we will obtain matrix S which consists of 8 singular values $\sigma_1, \sigma_2, \dots, \sigma_8$
3. Multiply S with scalar α and do round-off operation:

$$S_m = \text{floor}(\alpha S) \quad (3)$$

and to get binary bits, we use modular arithmetic as:

$$B = S_m \bmod 2 \quad (4)$$

So, we will obtain vector B that consists of 8 binary bits.

4. Tiling the binary bits into the size of image block, which is 8×8 . One alternative is by creating matrix B_r , where its row consists of vector B .
5. Permutate matrix B_r by random permutation based on seed key (Key-2).
6. Apply XOR operation between the scrambled binary bits (matrix B_r) and the watermark image block W_r to get the authentication data X_r :

$$X_r = B_r \oplus W_r \quad (5)$$

7. Embed the authentication data X_r to LSB of image block A_r to get the watermarked image block A_r^w . By doing this process for all blocks, we get the watermarked image.

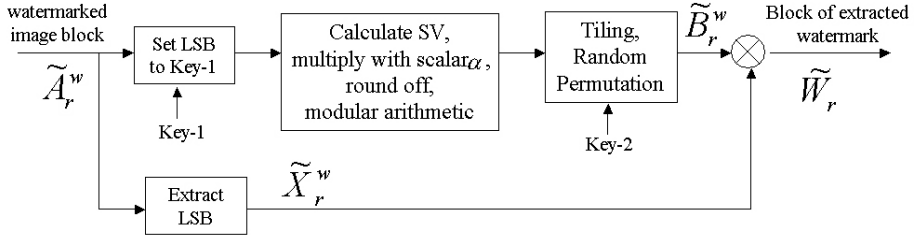


Fig. 2. Block diagram of watermark extraction process

4 Watermark Extraction

The extraction process is exactly the same as insertion process except that the last XOR operation is done between the scrambled binary bits B_r and the LSB value of watermarked image in block r .

The detailed process is as follow: first, divide the possibly tampered watermarked image we want to check into the 8×8 block size, as in embedding process. For each block, do the following steps:

1. Extract the LSB of watermarked image block \tilde{A}_r^w (we call it \tilde{X}_r^w).
2. Replace the LSB of \tilde{A}_r^w with Key-1 we used in the embedding process.
3. Calculate singular values (SVs) of \tilde{A}_r^w . We will obtain matrix S' which consists of 8 singular values $\sigma_1, \sigma_2, \dots, \sigma_8$
4. Multiply S' with scalar α that we used in embedding process and do round-off operation:

$$S'_m = \text{floor}(\alpha S') \quad (6)$$

and to get binary bits, we use modular arithmetic as:

$$B' = S'_m \bmod 2 \quad (7)$$

So, we will obtain vector B' which consists of 8 binary bits.

5. Tiling the binary bits into the size of image block, which is 8×8 as in the embedding process. We will get matrix \tilde{B}_r^w , where its row consists of vector B' .
6. Permutate matrix \tilde{B}_r^w using the same random permutation we used in the embedding process.
7. Apply XOR operation between the matrix \tilde{B}_r^w with the LSB of watermark image block \tilde{X}_r^w to get the extracted watermark image

$$\tilde{W}_r = \tilde{B}_r^w \oplus \tilde{X}_r^w \quad (8)$$

Note that the index w is used to indicate that the image block processed is the possibly watermarked image. By doing this process to whole block of watermarked image, we will get the extracted watermark image which from it we can tell whether the image is tampered or not. If the image is not tampered and the right keys are used, the result will be the right watermark image, clear without noise.



Fig. 3. Original image (*left*) and watermarked image (*right*)

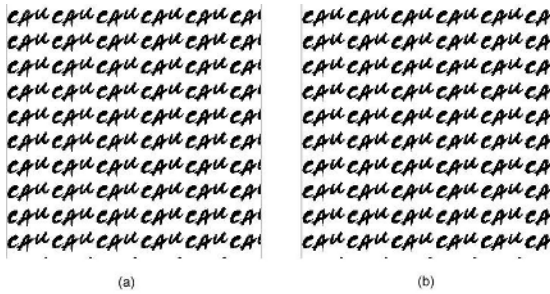


Fig. 4. (a)Original watermark image (b)Extracted watermark image when no tamper occurred

5 Experimental Results

Figure 3 shows original Couple image, and its watermarked image. No degradation in quality or any noticeable distortion. The PSNR is 54 dB. We use 512×512 Couple image, with scalar $\alpha=1000$ and 8×8 block size. Other scalar value and block size can be used. In general, higher scaling value is preferable to increase the sensitivity of SVs.

Figure 4(a) shows the watermark image we used in the experiment, and figure 4(b) is the extracted watermark image, retrieved from the untampered watermarked image, where the correct keys were used in the extraction process.

Figure 5 shows two versions of tampered images. Supposed that someone was intentionally erased the telephone on the desk (left) and added a little boy in the middle of this couple (right).

To detect whether the image is authentic or not, we extracted the watermark using the extraction process. As shown in figure 6 and 7, we can detect the tampered location for both cases. The resemble noise show specific area where the changes has been made to the watermarked image. So in both case the image is regarded as unauthentic.



Fig. 5. Tampered Images

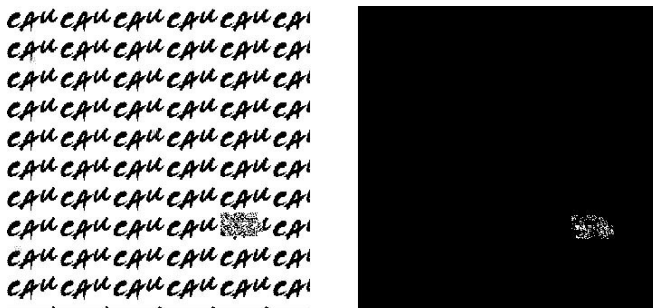


Fig. 6. Extracted watermark from the left tampered image in figure 5

If the wrong keys were used, the extracted watermark will be an image, which resembles noise, as shown in figure 8. The left image in figure 8 shows the extracted image if 1 bit of Key-1 was changed, and the right image in figure 8 shows the extracted image if different random permutation (Key-2) was used.

We also conducted a series of experiment for other image processing such as compression, scaling, and filtering, and the extracted watermark also resembled noise similar to the picture shown in figure 8. Only unmodified watermarked image produce a right watermark image and can be regarded as authentic.

6 Extension Algorithm for Color Images

In color image, we have three color channels R (red), G (green), and B (blue). We can apply this technique for color images, by embedding the watermark for each channel in color image, R, G, and B. We can embed independently watermark in each channel, and we can detect any tamper to each color channel. Or else, we can use XOR operation of the extracted watermarks to obtain only one watermark.

Below is the experimental result for color image using classic image 'baboon'. The tampered version of baboon image is shown in figure 9 where the baboon

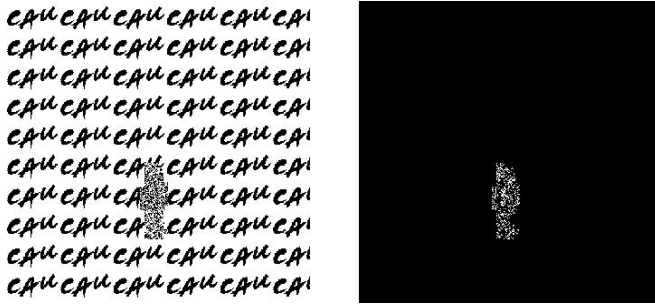


Fig. 7. Extracted watermark from the right tampered image in figure 5

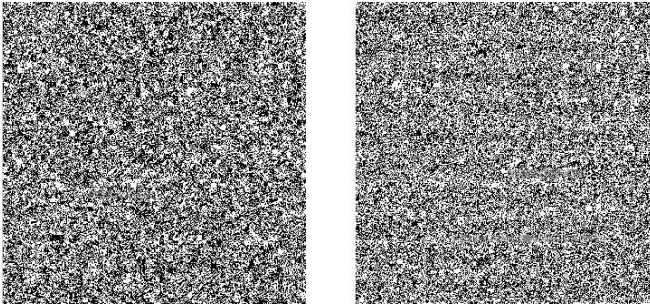


Fig. 8. Extracted watermark if the wrong keys were used



Fig. 9. (a) Tampered Baboon image (b) Extracted watermark image (c) Localization

eyes color is changed. From the extracted watermark image, we can tell that tampered has occurred and its location.

7 Conclusion

We propose a fragile watermarking technique utilizing singular values of the image block for image authentication. Using this technique we can check the authenticity of an image. The security of this technique relies on the two keys used in the embedding process. No cryptographic function is required. This technique also can detect location where the changes occurred.

In future work, we will extend the proposed scheme to authenticate JPEG images using quantized singular values of image.

Acknowledgement. The Ministry of Education, Seoul, Korea, supported this research under the BK21 project.

References

1. Yeung, M., Mintzer, F.: An Invisible Watermarking Technique for Image Verification. Proc. ICIP, vol. 2. Santa Barbara, CA (1997) 680-683
2. Memon, N., Shende, S., Wong, P.: On the Security of the Yeung-Mintzer Authentication Watermark. Proc. Of the IS and TPICS Symposium, Savannah, Georgia (2000)
3. Wong, P., Memon, N.: Secret and Public Key Image Watermarking Schemes for Image Authentication and Ownership Verification. IEEE Trans. on Image Processing, vol. 10 (2001) 1593-1601
4. Byun, S., Lee, S., Tewfik, A., Ahn, B.: A SVD-Based Fragile Watermarking Scheme for Image Authentication. International Workshop on Digital Watermarking. Lecture Notes in Computer Science vol. 2613 (2002) 170-178
5. Liu, R., Tan, T.: An SVD-based Watermarking Scheme for Protecting Rightful Ownership. IEEE Trans. on Multimedia, vol. 4 (2002) 121-128
6. Andrews, H., Patterson C.: Singular Value Decomposition (SVD) Image Coding. IEEE Trans. Commun., vol. COM-24 (1976) 425-432

Visual Cryptography for Digital Watermarking in Still Images

Gwo-Chin Tai^{1,2} and Long-Wen Chang¹

¹ Institute of Information Systems and Applications

National Tsing Hua University

101 Sections 2, Kung Fu Road

Hsinchu, Taiwan, 30055

² Chunghwa Telecom Labs

Yang-Mei, Taoyuan, Taiwan 326

maxwell@cht.com.tw

Abstract. Visual cryptograph can represent the secret image by several different shares of binary images. It is hard to perceive any clues about a secret image from individual shares. The secret message is revealed when parts or all of these shares are aligned and stacked together. In this paper, a novel public robust digital watermarking scheme based on visual cryptography is proposed. A binary logo is used to represent the ownership of the host image. The logo is used to generate a private sharing image and a public sharing image by visual cryptography algorithms. We use the public sharing image as the watermark embedded in the host image. An error correction-coding scheme is also used to protect the watermark. Simulation shows that our proposed watermarking algorithm is robust against various attacks indeed. Our algorithm and public share image can be open to the public, but only the owner, who uses the private sharing image of the ownership logo, can retrieve the logo for ownership. If there is an argument between the owner and the attacker the private sharing image can be provided to the arbitrator to resolve the ownership issue more convincingly.

Keywords: Visual cryptography, digital watermarking, wavelet packet transform.

1 Introduction

Visual cryptography [1][2] applies human vision to protect the secret message, which can be a text or an image. It represents the secret message by several different shares of binary images. It is hard to perceive any clues about a secret image from individual shares. However, parts or all of these shares are aligned and stacked together; the secret message will be revealed.

Digital watermarks have been proposed recently as the means for ownership protection of multimedia data. Xia [3] proposed a method to embed the watermark into the wavelet coefficients from high frequency subbands to low frequency subbands. The method [4] searches perceptually significant wavelet coefficients

















Pixel P	S_1	S_2	$S_1 + S_2$	Contrast
				50%
				50%
				100%
				100%

Fig. 1. Sharing and stacking scheme of black and white pixels.

for effective digital watermarking. Recently, some researcher proposed some digital watermarking algorithms based on the data encryption algorithm to enhance the authentication issues [5].

In this paper, a novel watermarking scheme with visual cryptography algorithm is presented. We employ the visual cryptography algorithm to generate two sharing images, the private share and the public share, from the ownership logo image. The public share is embedded in the cover image as a watermark. We also use an error correction-coding scheme to protect the watermark for various attacks. If we appropriately overlap the private share image and the extracted public share image from the watermarked image, we will recover the visual logo image. The private share image can be provided to the arbitrator when an argument of the ownership issue occurs.

2 Review of Visual Cryptography

Naor and Shamir proposed visual cryptography in 1995. In visual cryptography the secret image is decomposed into several sharing images. By overlapping the sharing images directly, one can find out the secret with eyes. In a t -out-of- n scheme, there would be n share images, and if any m , $m \geq t$ share of them are superimposed; the secret image should magically appear. However, examination of any m , $m \leq t$, shares should reveal no information about the secret image. In the simplest 2-out-of-2 visual threshold problem [1], a pixel W of the secret image is expanded to form two 2×2 blocks S_1 and S_2 , as shown in Fig. 1 on each of the two shares.

S_1 and S_2 contain one half of black and one half of white. Therefore, we cannot learn that P is white or black by just seeing S_1 and S_2 . If P is the black, randomly choose one of last two rows of Fig. 1 and the overlapped block $S_1 + S_2$ is black. If P is white, randomly choose one of upper two rows of Fig. 1 and half

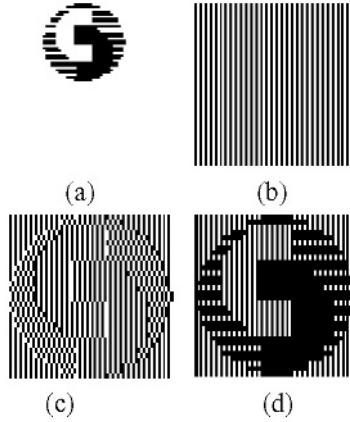


Fig. 2. Visual cryptography for Chunghwa Telecom logo (a) Secret image; (b) Share image 1; (c) Share image 2; (d) Recovering image.

of overlapped block $S_1 + S_2$ are black and the remaining half is black. Though P has contrast loss of 50% it can still be distinguished as white. Take Fig. 2 for example. The secret image (a) with the Chunghwa Telecom logo image is decomposed into two share images (b) and (c). When stacking the two share images, we can obtain the reconstructed image (d). Even though the contrast of the resulting image is degraded by 50% , human eyes can still identify the content of the secret image easily.

3 The Proposed Visual Cryptography for Digital Watermarking

We refer to the public key infrastructure based message identification model to establish the visual cryptography based message identification model [6]. Fig. 3 shows the whole procedure of share image generator [8]. The host image information is associated with other relevant information from Third Trusted Party (TTP), owner and the features of the cover image. Selecting an ownership logo image is as the secret image. The secret image and host image information such as host image identification are fed into 2-out-of-2 visual cryptography threshold schemes to producing share images.

We generate a pair of sharing images from the secret image by a 2-out-of-2 visual cryptography threshold scheme to achieve the message encryption. Take the public sharing image as the cipher messages to transmit. Only using the private sharing image to superimpose the public sharing, one can retrieve the visual image for the original message. The private sharing can be used to verify the identity of the message owner.

Suppose we turn our attention to a pixel P_s in the share S_1 . One of the two sharing blocks subpixels in P_s is black and the other is white. Furthermore, each

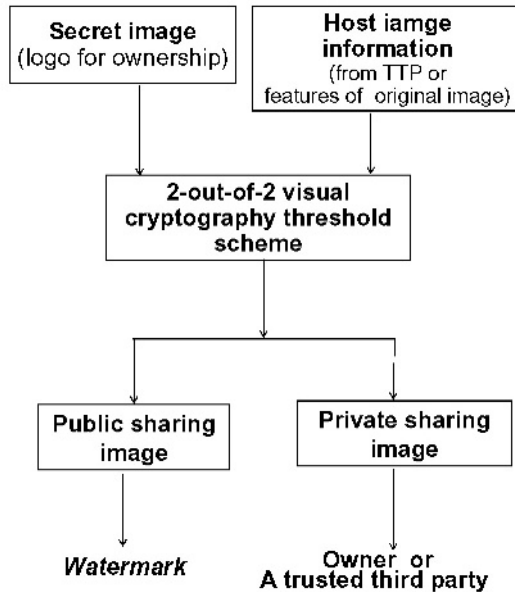


Fig. 3. The procedure of share image generator

of the two possibilities black-white and white-black is equally likely to occur, independent of whether the corresponding pixel in the secret image is black or white. Hence the share S_1 gives no clue as to whether the pixel is black or white. The same concept applies to the share S_2 . Since all the pixels in the secret image were encrypted using independent random "coin flip", there is no information to be gained by looking at any group of pixels on a share, either. Unless by an exhaustive method, we cannot find out the secret image for all possible situations of each picture element what is black or white.

In the traditional public key cryptography solution, if there are any malicious modifications on the content of the cipher message, the message cannot be retrieved even only change 1 bit. The recovered message will be random data. However, in the visual cryptography solution, if there are any malicious modifications on the content of the cipher message, the public share image, and the superimposition by two shares will result in a noise-like visual appearance. The decoded image is still visible.

Fig. 4 and Fig. 5 show the results of the recovered image when the share image 1 or share image 2 in Fig. 2 is corrupted by noise. We find that sharing image 1 is better than the sharing image 2 to resist against the noise corruption. Therefore, in simulation we choose sharing image 1 to be the watermark that embeds to the cover image for the proposed watermarking scheme. The proposed system has the following advantages:

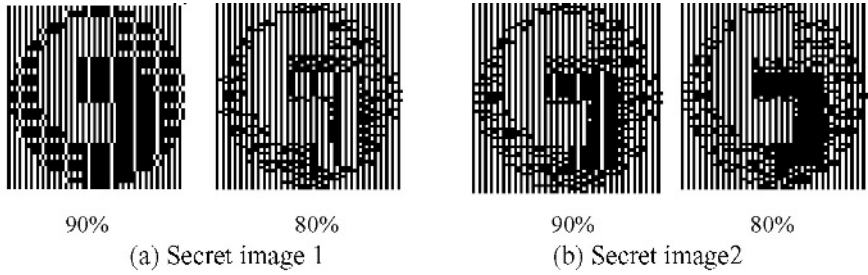


Fig. 4. Visual cryptography Recovering image for Chunghwa Telecom logo (a) Secret image 1; (b) the share image 2 after 4%, 6% noise corruption.

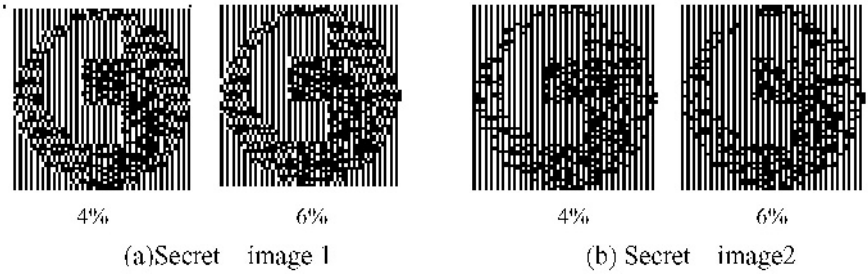


Fig. 5. Visual cryptography for Chunghwa Telecom logo (a) Secret image 1; (b) Share image 2 after 90%, 80% JPEG compression.

Security: By adopting content-based share image generator scheme (visual cryptography), the security of whole embedding and authentication procedure is guaranteed.

Robustness: The authentication is based on global visual effect. Any local defect due to noise will not affect the final decision.

Fig. 6 shows the block diagram of embedding the watermarks. We select the secret image W_o , a logo image for ownership, to form two sharing images, the share image pair named the private share image W_{Pr} and the public share image W_{Pk} , by share image generator. The W_{Pr} and W_{Pk} are random and unmeaning image. Inverse visual cryptography is to overlap the W_{Pr} and W_{Pk} to recover the original visual logo image W_{or} . The transform WPT denotes an orthogonal transform, wavelet packet transform, to decompose the original host image and find a list of frequency coefficients. Then, some significant coefficients are searched for embedding the watermark. The watermark W_{Pk} in the simulation is protected with error correcting code before it is embedded in these significant frequency coefficients. Finally, it is processed by the inverse wavelet packet transform WPT^{-1} to produce the watermarked image. After we search the significant wavelet coefficients we can embed the watermark [9] as

$$\omega_i' = \omega_i (1 + \alpha \times (-1)) \text{ If } I_p(i) = 0,$$

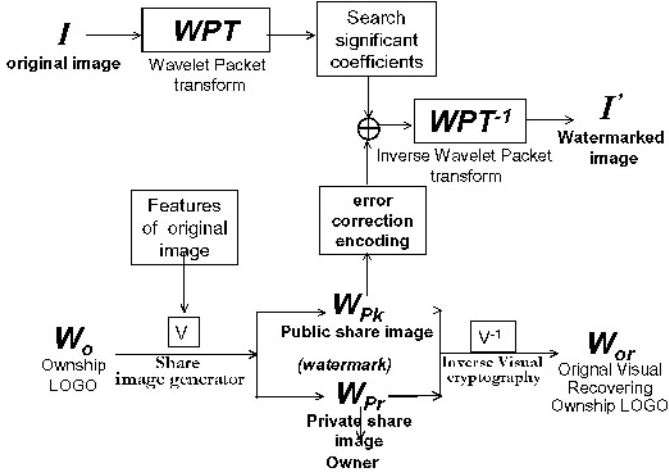


Fig. 6. The proposed public digital watermark embedding algorithm

$$= \omega_i (1 + \alpha \times (1)) \quad \text{otherwise;}$$

$I_p(i)$ is the watermark with error correcting. ω_i is the original frequency coefficient and is the embedded frequency coefficient and α is a scalar factor. The watermark extraction is the inverse of the watermark embedding. If there is an argument between the owner and the attacker the private share image can be provided to the arbitrator to reveals the visual recovering CHT logo W_{vr} after extracting the public share image from the watermarked image. Consequently, the proposed watermarking scheme can be open and resolve the copyright issue more convincingly.

4 Simulation Results

In the simulation, "Lena" is used as a test image. It is a gray levels image of 512×512 pixels. We use 2-out-of-2 visual threshold schemes to generate two sharing images, of 64×64 pixels from the bi-level CHT logo image of 32×32 pixels. The Hamming code is used as the error correction code. After embedding the watermark, we utilize the peak signal-to-noise ratio (*PSNR*) to consider the quality of watermarked image. We compute normalized correlation *NC* to decide whether watermark exists. If the correlation is larger than the predefined threshold, we assume watermark exists. If the extracted watermark logo is invisible, we can compare the *NC* value with other random watermarks to prove the existence of watermark [10]. Fig. 7 shows an example of the original image and its watermarked image by the proposed algorithm. Fig. 8 lists the simulation results under uniform noise corruption for various noise levels. Note that the extracted public share images are overlapped with the private share image to show the secret logo.



Fig. 7. (a) the original image (b) the watermarked image where PSNR=42.9423

	10%	20%	30%	40%
Extracted watermark				
Visual recovering logo				

Fig. 8. noise corruption; the noise are 2%, 4%, 6% and 8%, the PSNR are 27.28, 23.79, 20.89, and 19.99dB respectively

5 Conclusion

In this paper, we proposed a public digital watermarking with visual cryptography. According to our simulations, the public share can be extracted even the watermarked image are under various attacks. Our watermarking scheme in the simulation survives under attacks like noise and cropping operation. It is also strongly robust against JPEG compression and EZW compression. We utilize visual cryptography to protect the ownership logo and provide authentication. We also use error-correction code to improve the reliability of the extracted watermark. Therefore, our digital watermarking scheme is very robust and secure for image authentication. It can be open to the public because anyone cannot retrieval the logo without the private share image. The private share image can be provided to the arbitrator to resolve the copyright issue more convincingly.

References

1. M. Naor, A. Shamir: Visual cryptography, Lecture Notes in Computer Science, vol. 950 (1995) 1–12
2. M. Naor, A. Shamir: Visual cryptography II, Improving the contrast via, the cover base, in Security Protocols Work-shop (1996) 197–202
3. C. G. B. Xiang, Gen Xia and G. R. Arce: A Multiresolution Watermark for Digital Images, In 1997 International Conference on Image Processing (ICIP 97), (Santa Barbara, CA), IEEE Signal Processing Society, July (1997).
4. M. D. Swanson, B. Zhu, A.H. Tewfik: Multiresolution Scene-Based Video Watermarking Using Perceptual Models, IEEE Journal on Selected Areas in Comm., special issue on Copyright and Privacy Protection, vol. 16, no. 4, (May 1996) 540–555
5. Yong-Cong Chen and Long-Wen Chang: A Secure and Robust Digital Watermarking Technique by the block cipher RC6 and Secure Hash Algorithm, IEEE (2001) 518–521
6. M.Naor and Benny Pinkas: Visual Authentication and identification, in Advances in Cryptology-CRYPTO '97, B.Kaliski Jr. Ed., Vol. 1294 of “Lecture Notes in Computer Science” Springer-Verlag, Berlin, (1997) 322–336
7. Doug Stinson: Visual cryptography and threshold scheme, IEEE Potential, (Feb/Mar 1998) 13–19
8. Q.B. Sun, P.R. Feng and R. Deng: An Optical Watermarking Solution for Authenticating Printed Documents, IEEE Potential (2001) 66–70
9. I. Cox, J. Kilian, F. T. Leighton, and T. Shamoon: Secure Spread Spectrum Watermarking for Multimedia, Technical Report 95-10, NEC Research Institute, (1995)
10. Young-Chang Hou, Pei-Min Chen: asymmetric watermarking scheme based on visual cryptography, Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on, Volume: 2, 21-25 (Aug. 2000) 992–995
11. Zhi Zhou; Arce, G.R.; Di Crescenzo, G. : Halftone visual cryptography, Image Processing 2003. Proceedings. 2003 International Conference on, Volume: 1, Sept. 14-17 (2003) 521–524
12. Stinson D.R.: An introduction to visual cryptography, presented at Public Key Solutions '97. Available at <http://bibd.unl.edu/~stinson/VCS-PKS.ps>.

A New Object-Based Image Watermarking Robust to Geometrical Attacks

Jung-Soo Lee^{1,2} and Whoi-Yul Kim²

¹ MarkAny Inc., 10F, Ssanglim Bldg., 151-11, Ssanglim-dong,
Jung-gu, Seoul, 100-400, Korea

jslee@markany.com

² Dept. of Electrical and Computer Eng., Hanyang University,
17, Haengdang-dong, Seongdong-gu, Seoul, 133-791, Korea

{jslee, wykim}@vision.hanyang.ac.kr

Abstract. This paper presents a new approach for digital image watermarking providing robustness to geometrical distortions. As embedding watermark by the object, we can detect the embedded watermark having no concern with geometrical attacks. After selecting the area of object, we scan it by the direction of the major axis. Because the size of watermark inserted is always fixed we can detect the embedded watermark by adjusting the size of the selected object to that of the original watermark. Experimental results show that the proposed object-based watermarking algorithm invariantly detects the embedded watermark from the watermarked images.

1 Introduction

During the last decade, digital image watermarking techniques have grown dramatically. They had provided the robustness to the image compression and filtering attacks [1,2,3]. These attacks degrade the quality of images but do not change the original position of pixels. Because geometrical attacks change it, however, they prevent previous watermarking techniques from detecting the watermark. Therefore many researches with watermarking techniques to geometrical attacks have recently proceeded. But they did not provide an efficient way.

In this paper, we propose an object-based watermarking technique providing the robustness to geometrical attacks. It gives the efficient way to embed and detect watermark by processing in an object-base.

Previous Work

Pereia and Pun proposed a method that embeds a template with watermark into an image to deal with geometrical attacks [6]. This template is inserted into middle frequencies of the image spectrum and creates the local peaks. As finding the local peaks, we can synchronize the watermarked image with an original watermark. But this method has a disadvantage that it is difficult to find the local peaks because they move when the geometrical attacks happen to the watermarked image.

Patrick Bas *et al.* composed a geometrically robust watermarking system using feature points of an image [7]. Authors embed a triangular watermark into a triangular image selected from the connection of the three feature points. Because the feature points are moved or disappeared if geometrical attacks are applied to the watermarked image, however, it is difficult to extract the feature points identical to those of an original image and to make triangle composed of the identical three feature points to the original image.

The outline of this paper is as follows. Section 2 describes the method of object selection. Section 3 explains a watermark-embedding scheme. And the detection method of watermark embedded into images is described in section 4. Section 5 experiments how robust the proposed algorithm is to geometrical attacks. Finally we give conclusions on the proposed object-based watermarking technology and discuss the direction of the future researches.

2 Extracting the Area of Object

To embed the watermark by the objects, firstly we have to extract objects from an input image. In this section, we explain the method to select the area of objects. The way to extract them is as follows.

- a. Divide an image into objects through the labeling scheme like as fig.1 below.
- b. Find the edge of objects from the labeled image. As calculating the distance between two edge points repeatedly, we can find the major axis of the object.
- c. Calculate points having the longest distance from the major axis in the above and below part of it respectively. If a simple equation is ‘ $ax+by+c=0$ ’, the distance from point (x_1, y_1) to a straight line above is calculated as following.

$$D = \frac{|ax_1 + by_1 + c|}{\sqrt{a^2 + b^2}} \tag{1}$$

And then find the parallel straight lines with the major axis, which pass through two points found above.

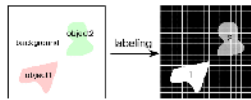


Fig. 1. The labeling of an image.

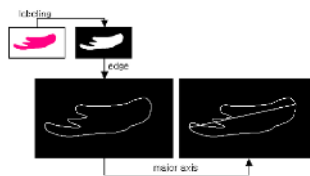


Fig. 2. Finding the major axis.

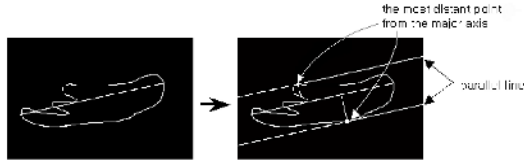


Fig. 3. The longest point.

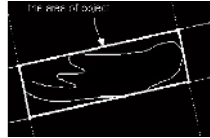


Fig. 4. The area of the object.

d. And then find two lines having right angles to the major axis, which pass through two points at the end of the major axis respectively.

Through the process above, when the area of object is decided from an image, we have to scan its area in the direction of the major axis, which is a rule to have the proposed algorithm be robust to geometrical attacks.

3 Watermark Embedding

In this section, we explain the watermark embedding method. The generation of a watermark is described in subsection 3.1. In subsection 3.2, the way to adjust the size of the generated watermark according to the selected object's area is explained. Finally, we embed the watermark generating from the process above.

3.1 Watermark Generation

User information that is composed of texts or Arabian numerals is firstly converted to binary type($\in\{0, 1\}$) and embedded in the spatial domain using shift control. And we use the modulation signal below to spread the binary sequence. This modulation signal is addressed as base watermark, which is generated by MATLABTM code.

```
rand('seed', 100);
baseWater = (round(rand(128,128))-0.5)×2
```

User information expressed by binary sequence is embedded to the object's area by combining this base watermark. That is, we express the binary sequence(user info.) by shifting the base watermark and adding each shift based watermark as fig. 5 below.

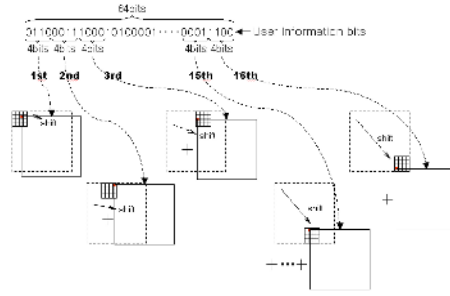


Fig. 5. Watermark generation(the shift and addition process of the base watermark to express user information).

The shift control is performed using eq. 2. That is to say, we have to shift the base watermark to the extent of S_x horizontally and S_y vertically according to 4bits' binary sequences.

$$\begin{aligned}
 S_x &= \{ \text{mod}_{(B_M/B_S)}(I_U) \} \times B_S + \frac{B_S}{2} + \{ \text{mod}_{(B_L/B_M)}(I_{U-nth} - 1) \} \times B_M \\
 S_y &= \left\lfloor \frac{I_U}{(B_M/B_S)} \right\rfloor \times B_S + \frac{B_S}{2} + \left\lfloor \frac{(I_{U-nth} - 1)}{(B_L/B_M)} \right\rfloor \times B_M
 \end{aligned}
 \tag{2}$$

Where, $I_U (\in \{0,1, \dots, 15\})$ means 4bits information. B_L , B_M and B_S indicate the size of large, middle and small block respectively and have 128, 32 and 8 respectively. And $\lfloor \bullet \rfloor$ means the integer doesn't exceed the result of operations. $\text{mod}_p(x)$ means the remainder resulted from dividing x into p . I_{U-nth} ($1^{st}, \dots, 16^{th}$) means where 4bits' data to be embedded locate(n th 4bits). For example, if $I_{U-nth} = 9$ and $I_U = 6$ ('01110' expressed in binary code), we should shift the base watermark to 20 pixels horizontally and 76 pixels vertically.

3.2 Watermark Embedding

To embed the generated watermark (in fig. 5) into the selected object's area, it has to be adjusted to the size of the selected object's area. And then we embed it into the selected object's area using the eq. 3.

Fig. 6 describes the watermark embedding process.

$$I_{ob}^{WM} = I_{ob} + \alpha \times I_{ob}^{hvs} \times WM_g
 \tag{3}$$

Here, I_{ob}^{WM} indicates the watermarked object and I_{ob} indicates the object image. α is constant as the control factor of watermark strength. And WM_g indicates the generated watermark in fig. 5.

I_{ob}^{hvs} is to consider the quality of the image to be watermarked. It is made by applying for the local strength of the watermark in the object. The local strength of the watermark is calculated through the following process.

- a. Apply the high pass filtering to the object's area of the image.
- b. Normalize the output above to an appropriate level.

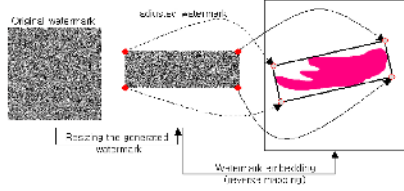


Fig. 6. Watermark embedding process.

4 Watermark Extracting

Watermark extraction process is analogous to watermark embedding process. First of all, after extracting the object we adjust the size of the object to that of the original base watermark. And then the extracted object is applied to high pass filter. It is why the watermark has to be embedded weakly for invisibility. And it strengthens the embedded watermark signal while removes the image signals. Using the high pass filter, we can efficiently remove the image signal because the energy of an image gets together to low frequencies.

Once because the synchronization base watermark with the extracted object through the extraction process of the object's area, information embedded can be obtained using the correlation between the object's area and the base watermark. It is able to obtain like as following MATLABTM code.

```
Corr = real(iff2(fft2(WM).*conj(fft2(OBJ))));
figure; mesh(Corr);
```

Here, WM is the original base watermark and OBJ is the selected object's area that is filtered by the high pass filter.

Using the code above, correlation peaks are obtained. And we can extract the embedded information using eq. 4 below from the correlation peaks.

$$I_{U-nth} = \left\lfloor \frac{y_{pp}}{B_M} \right\rfloor \times \left(\frac{B_L}{B_M} \right) + \left\lfloor \frac{x_{pp}}{B_M} \right\rfloor$$

$$E_I = \left(\left\lfloor \frac{\text{mod}_{B_M}(y_{pp})}{B_S} \right\rfloor \times \left(\frac{B_M}{B_S} \right) \right) + \left\lfloor \frac{\text{mod}_{B_M}(x_{pp})}{B_S} \right\rfloor \quad (4)$$

Where, E_I means the extracted information that is converted into binary code. And y_{pp} and x_{pp} indicate y-position and x-position of correlation peaks respectively. And I_{U-nth} means the position where the extracted 4bits' data are located.

5 Experimental Results

In this section, we test how robust the proposed algorithm is to geometrical attacks. We used the following sketched image(with sizing 600×600 pixel) as the test image and assumed that the background is black for convenience.

5.1 The Extraction of Object's Area

In this subsection, we examine whether the object is rightly extracted or not in geometrical attacks. Fig. 7 shows the extraction results of the object's area when the geometrical attacks happen to the watermarked image.

In case of fig. 7(b), the extracted object's area was rotated in a 180° comparing to the original one, which is caused by the rotation of object. And the rotation of object brings about the change of the scan direction. But the embedded watermark is successfully extracted because the watermark extraction process is performed on every 90° .

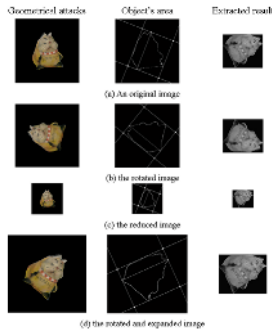


Fig. 7. The geometrically attacked image and the object's area extraction.

In (c)(d), the extracted object's size is changed. But we can extract the watermark through the adjusting process of object's size to the base watermark.

5.2 Watermark Extraction Results

In this subsection, watermark(user information) extraction results are shown. As seen in subsection 5.1, the extracted object is mostly invariant to geometrical attacks. But in case that the image was reduced seriously, we could not extract the watermark from the image because the loss of the embedded watermark was too serious to extract it. Table 1. shows the watermark extraction results. With the exception of the image shrunken seriously, we could verify that the embedded watermark is extracted successively in spite of attacks. In table 1, BER means bit error rate that is expressed in following formula.

$$BER = \frac{B_{Err}}{B_{Total}} \times 100(\%) \quad (5)$$

Where, B_{Err} means the error bits of embedded total bits and B_{Total} means the embedded total bits.

Table 1. Watermark extraction results.

Attacks	BER(%)
Rotation($1^\circ \sim 359^\circ$)	0
Resizing(30%~)	0
Resizing(< 30%)	45
Rotation & Resizing	0
Compression(>QF 25%)	0
Blurring, average	0
Sharpening	0

6 Conclusions and Future Works

In this paper, we proposed an object-based watermarking scheme that is robust to geometrical attacks. Because the direction of the object's scan is fixed to the major axis, our proposed algorithm is not affected to rotation attacks. In addition, as performing the process of adjusting the selected object's size to the original watermark, we make the proposed algorithm be invariant to scaling attacks. And because synchronization between the base watermark and the watermarked image is matched through the object extraction process, the embedded information is robustly extracted for other various attacks.

In the future, we will focus on the selection of the object. And we guarantee that will make our algorithm be more robust to various attacks than now.

Acknowledgement. Solideo Gloria. The present work has been supported by the NRL(National Research Institute) Project (2000N-NL-01-C-286).

References

1. I. Cox, J. Killian, T. Lighton, and T. Shamoon.: Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Processing*, vol. 6, Dec. (1997), 1673-1687.
2. B. Chen and G. W. Wornell.: An Information-theoretic approach to the design of robust digital watermarking systems. in *Proc. IEEE-ICASSP '99*, Phoenix, AZ, Mar. (1999).
3. D. Kundur and D. Hatzinakos.: Digital watermarking using multi-resolution wavelet decomposition. in *Proc. IEEE ICASSP '98*, vol. 5, Seattle, WA, May (1998), 2659-2662.
4. M. J. J. Meas and C. W. A. M. van Overveld.: Digital Watermarking by Geometric Warping. in *Proc. IEEE-ICIP '98*, vol. 2, Chicago, IL, Oct. (1998), 424-429.
5. M. Kutter.: Watermarking resisting to translation, rotation and scaling. *Proc. SPIE*, vol. 3528, Nov. (1998), 423-431.
6. S. Pereira and T. Pun.: Fast robust template matching for affine resistant image watermarking. in *International Workshop on Information Hiding*, ser. *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, vol. LNCS 1768, Sept. 29-Oct. 1,(1999), 200-210.
7. Patrick Bas, Jean-Marc Chassery, and Benoît Macq.: Geometrically Invariant Watermarking Using Feature Points. in *IEEE TRANSACTIONS ON IMAGE PROCESSING*, VOL. 11, NO. 9, Sept. (2002).

A Selective Image Encryption Scheme Based on JPEG2000 Codec

Shiguo Lian¹, Jinsheng Sun¹, Dengfeng Zhang², and Zhiquan Wang¹

¹ Department of Automation, Nanjing University of Science & Technology,
210094 Nanjing, China
sg_lian@163.com

² School of Mechanical Engineering, Nanjing University of Science & Technology,
210094 Nanjing, China

Abstract. In this paper, a novel image encryption scheme based on JPEG2000 is proposed, which encrypts some sensitive frequency subbands, bit-planes or encoding-passes selectively and partially. It is secure against such attack as known-plaintext attack or replacement attack. It is of low cost, keeps file format and compression ratio unchanged, supports direct bit-rate control, and does not degrade the original error-robustness. These properties make it suitable for real-time applications with direct bit-rate control requirement, such as web imaging, image communication, mobile or wireless multimedia, and so on.

1 Introduction

With the development of multimedia technology, the research on multimedia encryption becomes a hot topic. For the properties of large volumes and real-time requirement, multimedia data are difficult to be encrypted by traditional ciphers completely or directly. Therefore, better encryption algorithms are required.

Recently, a novel still image compression standard JPEG2000 [1] was announced and widely used, which make it necessary to study image encryption algorithms based on JPEG2000 codec. Till now, some algorithms have been reported [2-10], but they have some disadvantages that restrict their applications. For example, such permutation algorithms [2-5] are often of low cost, and keep file format unchanged. However, they are not secure enough against known-plaintext or select-plaintext attacks. Compared with them, complete-encryption algorithms [6,7], which encrypt compressed image data completely, can obtain high security. But they are often of high compute-complexity, and change the file format, which make them do not support such direct operations as image browsing, bit-rate control and so on. In practice, partial encryption algorithms [8-10] are more suitable for most applications since they obtain high speed by encrypting only some sensitive data. Wee [8] proposed a secure scalable streaming scheme for Motion-JPEG2000 codec, which supports direct bit-rate control. However, it encrypts most of the encoded data except some format headers, which restrict its applications in such fields as real-time transmission or mobile/wireless multimedia. Pommer [9] proposed a selective encryption scheme for wavelet-packet

encoded images, which is of low cost. But, it encrypts only tree structures while no coefficients' value, so the security cannot be confirmed for different images. Norcen [10] proposed a selective encryption scheme for JPEG2000 bitstream, which encrypts 20% of the compressed bitstream except format information. It is of low cost, and supports direct bit-rate control. However, the encryption scheme is not suitable for all the encoding modes.

Here, we propose an image encryption scheme based on JPEG2000 codec. This scheme encrypts bit-planes, code blocks or subbands selectively and partially according to the sensitivity to images' understandability. It is of low cost, keeps file format and compression ratio unchanged, supports direct bit-rate control, and is easy to be realized. The rest of the paper is organized as follows. In Section 2, the image encryption scheme is proposed. And its performances of security, compute-complexity, bit-rate control or error robustness are analyzed in Section 3. Finally, some conclusions are drawn, and future work is proposed in Section 4.

2 The Proposed Encryption Scheme

JPEG2000 codec is based on the embedded block coding with optimized truncation (EBCOT) scheme [11]. For JPEG2000 encoded images, we propose to encrypt them selectively. That is, only some parts of the code stream are encrypted while others are left unencrypted. The time-efficiency of this encryption scheme is obtained by reducing the data volumes to be encrypted. But its security depends on the selection of the parts to be encrypted and the selection of encryption algorithms. And it is constructed according to such principles: 1) The parts with high sensitivity prefer to be encrypted. Here, the parts' sensitivity means the effect on images' understandability. 2) The encrypted parts should be independent from the unencrypted ones. It is reasonable to understand that, the encrypted parts can be recovered from the unencrypted ones when there are some relationships between them. 3) The encryption algorithms with high security are preferred.

2.1 Sensitivity Test

Considering that images are encoded progressively in JPEG2000, we propose to select the suitable parts according to the subband, bit-plane or encoding-pass. In the following content, we test the sensitivity of the proposed three aspects (subband, bit-plane or code pass) through experiments.

Subband Sensitivity. In L-level wavelet transform, an image is transformed into $3L+1$ subbands. For example, the subbands after 5-level wavelet transform are LL5, LH5, HL5, HH5, . . . , LH1, HL1 and HH1. Each subband has different sensitivity to images' understandability. Taking Lena (256*256, gray, 5-level

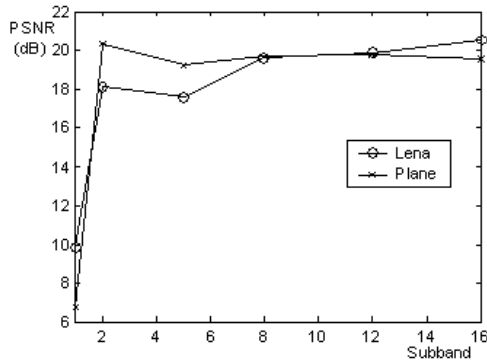


Fig. 1. The test of subband sensitivity. Where, x-axis represents the subbands that are numbered from the lowest frequency to the highest one.

wavelet transform) and Plane (512*512, gray, 5-level wavelet transform) for example, we encrypt each of the subbands, and compute the PSNR (Peak Signal-to-Noise Ratio) of the according image. The experimental results are shown in Fig. 1. As can be seen that, the images' quality is the worst when the lowest frequency (LL5) is encrypted. It tells that, the subband with the lowest frequency is of higher sensitivity to images' understandability, and it is prefer to be encrypted greatly.

Bit-plane Sensitivity. Similarly, for a transformed image, each bit-plane of the coefficients has different sensitivity to images' understandability. Taking the images proposed above for example, we encrypt each of the bit-plane, get the encrypted image, and show the sensitivity in Fig. 2. Seeing from the curves, the images' quality decreases with the rise of the number of the encrypted bit-plane. Additionally, the curves' gradient becomes smaller when the number of bit-plane is between 6 and 10. This indicates that it is better to encrypt all these bit-planes (6-th to 10-th). Thus, the significant bit-planes are more sensitive to images' understandability than less significant ones, which should be encrypted in order to achieve high security.

Encoding-pass Sensitivity. In bit-plane encoding, significant pass, refinement pass and cleanup pass have different sensitivity to images' understandability, which is tested here with the similar method proposed above. Fig. 3 is an example for Lena (256*256, gray, 5-level wavelet transform). The curves are all horizontal lines. That is because the JPEG2000 decoder can stop automatically when it detects some bit-errors. As can be seen that, the curve according to cleanup pass is the lowest one. Thus, the cleanup pass is more sensitive than other two passes under any bit-rate, and it is prefer to be encrypted.

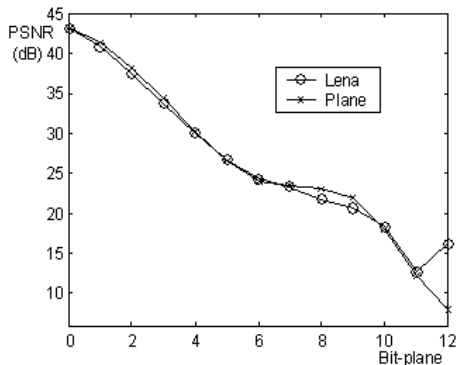


Fig. 2. The test of bit-plane sensitivity. Where, x-axis represents the bit-planes that are numbered from the least significant one to the most significant one.

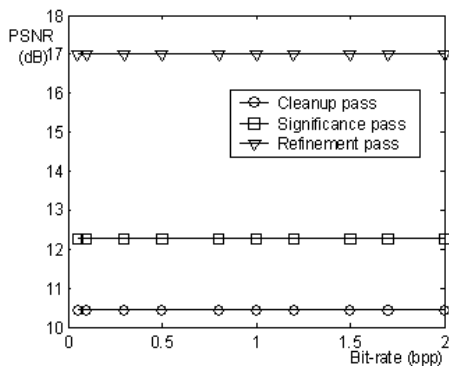


Fig. 3. The test of encoding-pass sensitivity. Where, x-axis represents the bit-rate (bits per pixel), and y-axis represents the PSNR of the encrypted image.

2.2 Independent Relationship Analysis

In the proposed three aspects (subband, bit-plane and encoding-pass), bit-planes are independent from each other, encoding-passes are also independent from each other, but subbands are not. In each code block, the bit-planes are independent from each other, so they can be encrypted selectively. For each bit-plane, the three encoding-passes are also independent from each other, which makes them can be selectively encrypted. In wavelet transform, the subbands in different layers depend on each other. That is, the subbands in the lower layer can be recovered from the ones in the higher layer. Therefore, it's not secure to leave all the subbands in a layer unencrypted.

2.3 Selection of Encryption Algorithms

The main decision to be made is whether to use stream ciphers or block ciphers, which depends on the extra performance of the encrypted data stream. For example, stream ciphers or the block ciphers with variable plaintext size are often selected if direct bit-rate control is required. If transmission errors cannot be avoided, then the bitstream should be encrypted segment by segment in order to reduce the effect caused by bit-errors. In JPEG2000 codec, since the encoding-passes are often of variable size, the block ciphers with variable size or stream ciphers are more suitable. In practice, stream ciphers are based on random sequences that can now still not be generated practically. Thus, the block ciphers with variable size [12] are more suitable here, such as AES, RC4, and so on.

2.4 The Proposed Encryption Process

According to the above analyses, we propose a secure encryption scheme. Where, the whole image is divided into two parts according to frequency range: low-frequency part and high-frequency part. And they are encrypted with different methods. Taking 3-level wavelet transform for example, the subband (LL3) in the third decomposition belongs to the low-frequency part, and the others belong to the high-frequency part. Thus, the encryption scheme is described as follows. 1) Low-frequency part. All the code blocks in the low-frequency part are encrypted. For each code block, only the significant bit-planes (6th to 10th) are encrypted. And for each bit-plane, the three passes are all encrypted. 2) High-frequency part. In high-frequency part, all the code blocks are encrypted. For each code block, only the significant bit-planes (6th to 10th) are encrypted. And for each bit-plane, only the cleanup pass is encrypted. 3) Cipher selection. AES algorithm [12] is adopted to encrypt the selected parts pass by pass.

3 Performance Analysis

3.1 Security

The proposed encryption scheme has high perception security that can be testified by the experimental results shown in Fig. 4.

The security against brute-force attack, statistic attack or differential attack is determined by the adopted cipher. As is different from text encryption that, knowing only few encoded data does little help to the whole image. The adopted block ciphers, such as AES or RC4, both have high security against statistic or differential attack, which keeps the image encryption system of high security. The encryption scheme is secure against select-plaintext attack [13] or replacement attack [14]. In multimedia encryption, select-plaintext attack is testified effective to break permutation algorithms, in which, coefficients are only permuted in position, while their values keep unchanged. Replacement attack means to replace some of the encrypted data with other ones in order to reconstruct the plaintext under the condition of ciphertext-only attack.

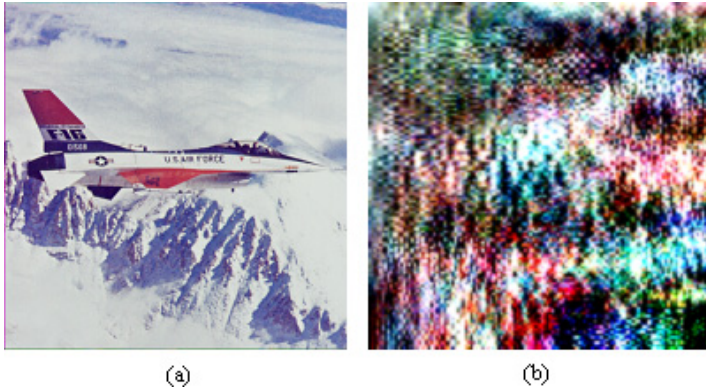


Fig. 4. The encryption results of image Plane (512*512, colorful, 6-level wavelet transform). Image (a) is the original one, and image (b) is the encrypted one. The encrypted image is too chaotic to be understood. It shows that, the proposed encryption algorithm is of high perception security.

3.2 Computational Complexity

The computational complexity of the proposed cryptosystem depends mainly on the bitstreams to be encrypted. It encrypts only few sensitive data while keeping high security, and thus obtains low cost. For various images, we test the encryption data ratio that is the ratio between the encrypted data and the whole bitstream, also the encryption time ratio that is the ratio between encryption time and encoding time. The results are shown in Table 1. Here, the encryption data ratio is no more than 20%, and the encryption time ratio is smaller than 15%. Thus, the encryption process does little effect on the encoding process, and the similar result is obtained on decryption process.

Table 1. Encryption speed test

Image	Size	Color	Encryption data ratio	Encryption time ratio
Lena	128*128	Gray	13.4%	8.9%
Boat	256*256	Gray	10.8%	10.6%
Cameraman	256*256	Gray	14.6%	8.7%
Village	512*512	Gray	16.8%	11.1%
Lena	128*128	Colorful	14.7%	9.4%
Jet	256*256	Colorful	15.2%	8.2%
Peppers	256*256	Colorful	12.1%	10.4%
Baboon	512*512	Colorful	15.3%	9.7%

3.3 Other Performances

In JPEG2000 codec, bit-rate control is realized by directly cutting some passes. The proposed encryption algorithm supports arbitrary size of plaintext, and supports direct ciphertext cutting from the end. This property makes it suitable for wider applications with direct bit-rate control requirement. Additionally, it encrypts each encoding-pass independently, which makes the data streams in different encoding-passes do not affect each other. Therefore, the error-passes cannot affect the previous passes, which does not reduce the original codec's robustness to transmission errors.

4 Conclusions and Future Work

In this paper, a JPEG2000 based image encryption scheme is proposed. It encrypts some wavelet coefficients selectively. Experimental results show that, the encryption scheme is secure against brute-force attack, select-plaintext attack or replacement attack. Additionally, it is time-efficient, does not change compression ratio, supports direct bit-rate control, and keeps the original error-robustness. These properties make it suitable for many applications such as web imaging or image transmission. What's more, it can be extended to MPEG4 codec, which is our future work.

Acknowledgement. This work was supported by the National Natural Science Foundation of China through the grant number 60374066 and the Province Natural Science Foundation of China through the grant number BK2004132.

References

1. ISO/IEC 15444-1: Information Technology-JPEG 2000 Image Coding System-Part 1: Core Coding System (2000)
2. Grosbois, R., Gerbelot, P., Ebrahimi, T.: Authentication And Access Control in the JPEG2000 Compressed Domain. In: Proc. SPIE 46th Annual Meeting, Applications of Digital Image Processing XXIV, Vol. 4472. San Diego (2001) 95–104
3. Ando, K., Watanabe, O., Kiya, H.: Partial-Scrambling of Still Images Based on JPEG2000. In: Proceeding of the International Conference on Information, Communications, and Signal Processing. Singapore (2001)
4. Ando, K., Watanabe, O., Kiya, H.: Partial-Scrambling of Images Encoded by JPEG2000. IEICE Trans., Vol. J85-D-11, No. 2 (2002) 282–290
5. Fukuhara, T., Ando, K., Watanabe, O., Kiya, H.: Partial-Scrambling of JPEG2000 Images for Security Applications. ISO/IEC JTC 1/SC29/WG1, N2430
6. Qiao, L., Nahrstedt, K.: A New Algorithm for MPEG Video Encryption. In: Proceeding of the First International Conference on Imaging Science, Systems and Technology (CISST'97). Las Vegas Nevada (1997) 21–29
7. Qiao, L., Nahrstedt, K.: Comparison of MPEG Encryption Algorithm. International Journal on Computers and Graphics. Vol. 22, No. 4 (1998) 437–448

8. Wee, S., Apostolopoulos, J.: Secure Scalable Streaming and Secure Transcoding with JPEG-2000. Technical Report, HPL-2003-117 (2003)
9. Pommer, A., Uhl, A.: Selective Encryption of Wavelet-Packet Encoded Image Data: Efficiency and Security. *Communications and Multimedia Security* (2003) 194–204
10. Norcen, R., Uhl, A.: Selective Encryption of the JPEG2000 Bitstream. *IFIP International Federation for Information Processing, LNCS 2828* (2003) 194–204
11. Taubman, D.: High Performance Scalable Image Compression with EBCOT. *IEEE Trans. on Image Processing*, Vol. 9, No. 7 (2000) 1158–1170
12. Buchmann, J.A. *Introduction to Cryptography*. Springer-Verlag New York (2001)
13. Chang, C.C., Yu, T.X.: Cryptanalysis of An Encryption Scheme for Binary Images. *Pattern Recognition Letters*, Vol. 23, No. 14 (2002) 1847–1852
14. Podesser, M., Schmidt, H.P., Uhl, A.: Selective Bitplane Encryption for Secure Transmission of Image Data in Mobile Environments. In: *CD-ROM Proceedings of the 5th IEEE Nordic Signal Processing Symposium (NORSIG 2002)*. Tromsø-Trondheim Norway (2002)

VQ-Based Gray Watermark Hiding Scheme and Genetic Index Assignment

Feng-Hsing Wang¹, Jeng-Shyang Pan², Lakhmi Jain¹, and
Hsiang-Cheh Huang³

¹ School of Electrical and Information Eng., University of South Australia
wanfy002@students.unisa.edu.au

² Dept. of Electronic Eng., National Kaohsiung University of Applied Sciences,
Kaohsiung, Taiwan, R. O. C.
jspan@cc.kuas.edu.tw

³ Dept. of Electronic Eng., National Chiao Tung University, Hsinchu, Taiwan,
R. O. C.
hchuang@mail.nctu.edu.tw

Abstract. A vector quantisation (VQ) based watermarking scheme for hiding the gray watermark is presented. It expands the watermark size, and employs VQ index assignment procedure with genetic algorithm, called genetic index assignment (GIA), for watermarking. The gray watermark is coded by VQ, and obtained indices are translated into a binary bitstream with a much smaller size. We partition the codebook into two sub-codebooks, and use either one of them based on the value of the bit for embedding. Next, GIA is employed to find better imperceptibility of watermarked image. Experimental results show that the proposed method possesses advantages over other related methods in literature.

1 Introduction

Digital watermarking techniques [1,2] generally can be classified into two categories: spatial-domain techniques and transform-domain techniques. Lately, more and more researchers have paid attention to the vector quantisation (VQ) [3] based watermarking schemes [4]–[7] since these schemes can enhance the conventional VQ system with the ability of watermarking.

For the VQ-based image watermarking schemes, Lu and Sun [4] proposed partitioning the VQ codebook into groups according to a secret key. They then adjusted the obtained VQ indices to hide watermark bits with these groups. Their method requires the original cover image to be presented during extraction. To improve this, Jo and Kim [5] introduced their method where they partitioned the used codebook into three sub-codebooks and used two of them to hide watermark bits. Wang *et al.* [6] presented another method, which is similar to the above methods, with a genetic codebook partition procedure. They employed the genetic algorithms (GA) [8] into the codebook partition procedure to obtain better performance. Huang *et al.* [7] introduced another type of watermarking schemes where they hid watermark bits into the output keys by referring to the

VQ indices. Therefore, their scheme provides better performance in quality and robustness than others.

In this paper, a VQ-based watermarking scheme adapted from [6] for hiding a gray watermark is presented. It applies the VQ coding procedure to compress the original watermark into smaller size and embeds the coded result into the cover image in the VQ domain. Furthermore, a genetic index assignment procedure which takes the concept of index assignment and the technique for optimization into account are proposed to improve the imperceptibility of the watermarking scheme.

2 VQ-Based Watermarking Scheme for Gray Watermark

2.1 Codebook Partition and Sub-codebook Extraction

For a codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_L\}$ used in the VQ system, we partition it into two sub-codebooks \mathbf{G}_0 and \mathbf{G}_1 according to a codebook partition key $\mathbf{K}_{\text{CP}} = \{g_1, g_2, \dots, g_L \mid \forall g_i \in [0, 1], i \in [1, L]\}$. Here $\mathcal{C} = \mathbf{G}_0 \cup \mathbf{G}_1$, $\mathbf{G}_0 \cap \mathbf{G}_1 = \emptyset$, and g_i of \mathbf{K}_{CP} assigns codeword \mathbf{c}_i from \mathcal{C} to sub-codebook \mathbf{G}_{g_i} . Then, we extract out another sub-codebook \mathcal{C}_S from \mathcal{C} by referring to another user-key \mathbf{K}_S . Here $\mathcal{C}_S \subset \mathcal{C}$ and \mathbf{K}_S defines which codewords in \mathcal{C} are selected to form \mathcal{C}_S . Sub-codebooks \mathbf{G}_0 and \mathbf{G}_1 will be used in the embedding procedure and sub-codebook \mathcal{C}_S will be used in the gray watermark coding procedure.

2.2 Gray Watermark Encoding and Decoding

For a gray watermark \mathbf{W}_G , we decompose it into N non-overlapping blocks and apply the conventional VQ encoding procedure with sub-codebook \mathcal{C}_S . The VQ indices of all the obtained nearest codewords are collected as index set $\mathbf{I}_W = \{y_1, y_2, \dots, y_N \mid \forall y_i \in [0, L_S], i \in [1, N]\}$, where L_S is the size of \mathcal{C}_S . We then translate \mathbf{I}_W into a binary bit stream \mathbf{W}_B according to L_S . For example, if $y_1 = 5$ and $L_S = 16$ (which means the length of each index is $\log_2(L_S) = 4$ bits), then y_1 is translated as $\{0,1,0,1\}$ because $(5)_{10} = (0101)_2$. The bits $\{0,1,0,1\}$ then form the first four bits of \mathbf{W}_B . The above process is repeated until all the indices in \mathbf{I}_W have been translated. We summarize the whole encoding procedure as Eq. (1) and the inverse procedure of it as Eq. (2). An example of the gray watermark coded result is given in Fig. 1.

$$\mathbf{W}_B = \text{Encode} (\mathbf{W}_G, \mathcal{C}_S) . \quad (1)$$

$$\mathbf{W}_G = \text{Decode} (\mathbf{W}_B, \mathcal{C}_S) . \quad (2)$$

2.3 Embedding and Extracting Procedures

To hide the watermark bits $\mathbf{W}_B = \{w_1, w_2, \dots, w_M\}$ into the cover image \mathbf{X} , which is decomposed into M non-overlapping blocks $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$ beforehand, the conventional VQ encoding procedure is performed to \mathbf{X} with \mathbf{G}_0 or

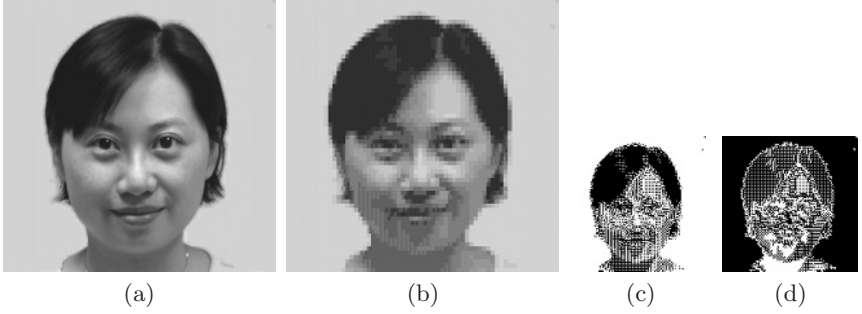


Fig. 1. An example of the proposed gray watermark encoding procedure: (a) the original gray watermark ($256 \times 256 \times 8$ bits/pixel), (b) the VQ encoded watermark ($256 \times 256 \times 8$ bits/pixel), (c) the binary bit stream, which is arranged as a binary image of size 128×128 pixels, obtained from the VQ indices of (b), and (d) the binary bit stream, which is also arranged as a binary image of size 128×128 pixels, obtained from the indices of (b) while assigning different indices to the codewords

\mathbf{G}_1 merely. That is, for a given block \mathbf{x}_i , $1 \leq i \leq M$, and the watermark bit w_i to be hidden in this block, sub-codebook \mathbf{G}_{w_i} is regarded as the used codebook in the VQ encoding procedure. From \mathbf{G}_{w_i} , a nearest codeword to \mathbf{x}_i can be obtained and outputted as the watermarked block \mathbf{x}'_i . The above process is repeated until all the watermark bits are embedded into the corresponding input blocks. After that, a watermarked image \mathbf{X}' is obtained by piecing together all the watermarked blocks $\{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_M\}$.

To extract the hidden watermark bits from a suspicious image $\hat{\mathbf{X}}$, we first segment $\hat{\mathbf{X}}$ into M non-overlapping blocks $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_M\}$. Then, we merely carry out the conventional VQ encoding procedure with the original codebook \mathbf{C} to obtain the nearest codewords for $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_M\}$. After that, the groups where the obtained codewords belong can be determined by referring to the user-key \mathbf{K}_{CP} . That is, for an input block $\hat{\mathbf{x}}_i$, $1 \leq i \leq M$, if the obtained nearest codeword to it is \mathbf{c}_j , $1 \leq j \leq L$, the hidden watermark bit w'_i will be $\mathbf{K}_{CP}[j] = g_j$. By piecing together all the extracted bits, a binary bit stream \mathbf{W}'_B is formed. Finally, the original gray watermark \mathbf{W}'_G can be recovered using Eq. (2).

3 Genetic Index Assignment (GIA)

In this section, we show the technique to modify the signal of the original watermark and the procedure to obtain a better index assignment key for better watermarking results. The concept of index assignment and GA [8] are employed respectively.

3.1 Index Assignment

In Sect. 2.2, sub-codebook $\mathbf{C}_S = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{L_S}\}$ is used to encode the gray watermark \mathbf{W}_G . When no index assignment introduced, the original indices of

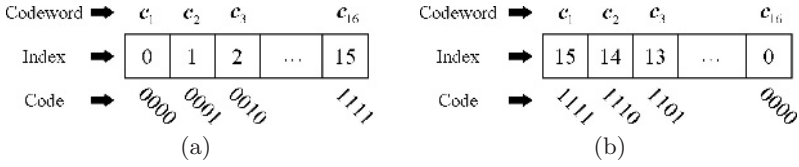


Fig. 2. Assigning different indices to the codewords: (a) the original assigned indices and (b) the reassigned indices

codewords $\{c_1, c_2, \dots, c_{L_S}\}$ are $\{0, 1, \dots, L_S - 1\}$ respectively. If we assign new indices to all the codewords in C_S as $\{L_S - 1, L_S - 2, \dots, 0\}$, for example, then the binary bit stream generated from W_G by applying Eq. (1) will be different from the original one. An example of assigning different indices to the codewords is shown in Fig. 2, where we assume $L_S = 16$ and the obtained nearest codeword is c_2 for a given block of W_G . When no index assignment introduced, the index of c_2 , which is 1 (see Fig. 2(a)), is translated as $\{0, 0, 0, 1\}$ by using the method introduced in Sect. 2.2. After assigning new indices $\{L_S - 1, L_S - 2, \dots, 0\}$ to the codewords, as Fig. 2(b), the index of c_2 now is 14, thus it is translated as $\{1, 1, 1, 0\}$. Fig. 1 (d) gives the coded result when index assignment introduced.

Based on the concept above, the watermarking scheme mentioned in Sect. 2 is modified as Fig. 3, where the index assignment key, K_{IA} , is designed for assigning new indices to the codewords in C_S , e.g., $K_{IA} = \{0, 1, \dots, 15\}$ in the above first example and $K_{IA} = \{15, 14, \dots, 0\}$ in the second example.

3.2 The GIA Procedure

We define a GA chromosome P as a set of L_S positive integers $\{y_1, y_2, \dots, y_{L_S}\}$, where $\forall y_i \in [0, L_S)$ and $y_i \neq y_j$ if $i \neq j$. Then, we apply the VQ encoding procedure to the gray watermark W_G , and collect the indices of the obtained nearest codewords as set I_W (see Sect. 2.2). Afterwards, the steps below are executed:

1. Generate m GA chromosomes $\{P_1, P_2, \dots, P_m\}$ randomly.
2. Convert I_W as $\{I_{IA_1}, I_{IA_2}, \dots, I_{IA_m}\}$ by applying the method mentioned in Sect. 3.1 with $\{P_1, P_2, \dots, P_m\}$ respectively. Here P_* is seen as K_{IA} .
3. Translate $\{I_{IA_1}, I_{IA_2}, \dots, I_{IA_m}\}$ into binary strings $\{W_{B_1}, W_{B_2}, \dots, W_{B_m}\}$.
4. Embed $\{W_{B_1}, W_{B_2}, \dots, W_{B_m}\}$ into X respectively to obtain m watermarked images $\{X'_1, X'_2, \dots, X'_m\}$.
5. Calculate the fitness scores $\{f_1, f_2, \dots, f_m\}$ for all the watermarked images respectively. Here due to the motivation of the GIA procedure is to obtain a better K_{IA} for better watermarking results in quality, therefore we borrow the well-known peak-signal-to-noise-ratio (PSNR) as the fitness function.
6. Select n ($n \leq m$) chromosomes with best fitness scores from current generation and from the best-list (for the first generation, there is no chromosome recorded in this list). The n chromosomes are then seen as the new best-list.

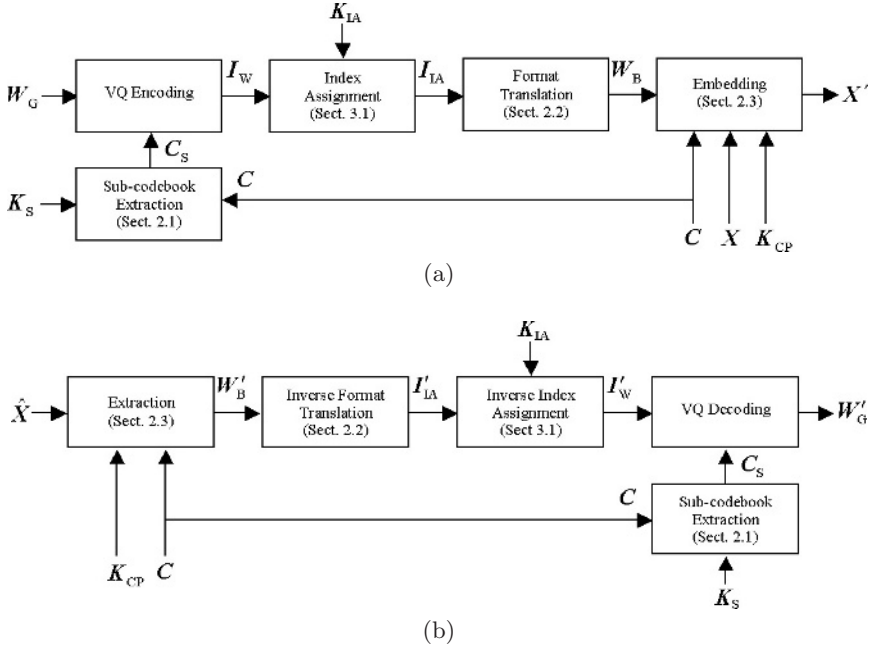


Fig. 3. The VQ-based watermarking scheme with a key for index assignment: (a) embedding and (b) extraction

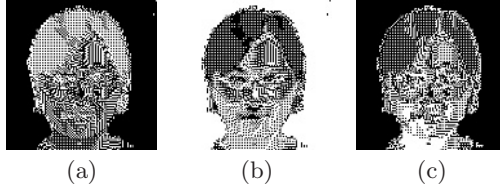
7. Terminate the training procedure and use the chromosome with the best fitness score as K_{IA} if the considered GA iteration t is met.
8. Regenerate m chromosomes for next generation by using the chromosomes recorded in the best-list. The better fitness score the chromosome has, the higher probability the chromosome will have to be selected as the parent.
9. Mutate the content of the m chromosomes with a mutation rate ρ (e.g., exchanging the values of two genes randomly).
10. Go to Step 2.

4 Simulation Results

In our simulation, the well-known images, LENA, PEPPERS, and BABOON, were used as the cover images. The size of either the cover image is 512×512 pixels in gray-level. The image shown in Fig. 1(a) with size 256×256 pixels in gray-level was used as the original watermark. A codebook which contains 256 codewords was obtained from the image of LENA by applying the LBG algorithm [3] with a threshold of 0.0001. We used it to encode all the three cover images. The gray watermark and either the cover image were decomposed into $N=4096$ and $M=16384$ non-overlapping blocks of size 4×4 pixels respectively. We used a random-generated key to partition the codebook. The extracted sub-

Table 1. Embedding results (PSNR) of the proposed methods (unit: dB)

Methods	LENA	PEPPERS	BABOON
VQ Coding	31.800	28.028	22.537
No GIA	29.504	26.775	22.060
With GIA	29.920	27.142	22.120

**Fig. 4.** The binary bit streams (which are arranged as binary images of size 128×128 pixels for display) obtained from Fig. 1(a) when the GIA procedure introduced: (a) LENA was used as the cover image, (b) PEPPERS was used as the cover image, and (c) BABOON was used as the cover image**Table 2.** Robustness test of the proposed methods (unit: dB)

Attacks	LENA	PEPPERS	BABOON
VQ, $L=512$	28.115	24.006	26.399
JPEG, QF=60%	20.667	17.944	19.257
JPEG, QF=80%	29.855	27.173	28.495
Median Filtering	15.699	15.163	11.659
Cropping, 25%	16.960	16.960	16.960

codebook for coding the gray watermark contains 16 codewords therein. The settings for the GIA procedure are: $m = 20$, $t = 1000$, $n = 10$, and $\rho = 30\%$.

Table 1 lists the embedding quality of the proposed methods. Fig. 4 displays the encoded results of the gray watermark when the GIA procedure introduced. Table 2 lists the robustness against some common image processing procedures (where QF denotes “quality factor”). Fig. 5 shows the recovered gray watermarks extracted from the attacked watermarked images listed in Table 2 while the image of LENA was used as the cover image and the GIA procedure introduced.

To compare with other results in literature, we also studied some related VQ-based watermarking methods in literature and carried out the simulations. To make the comparison objectively, Fig. 1(c) was used as the input binary watermark for those schemes that can only embed a binary watermark into the cover image, the image of LENA was used as the cover image, and other materials

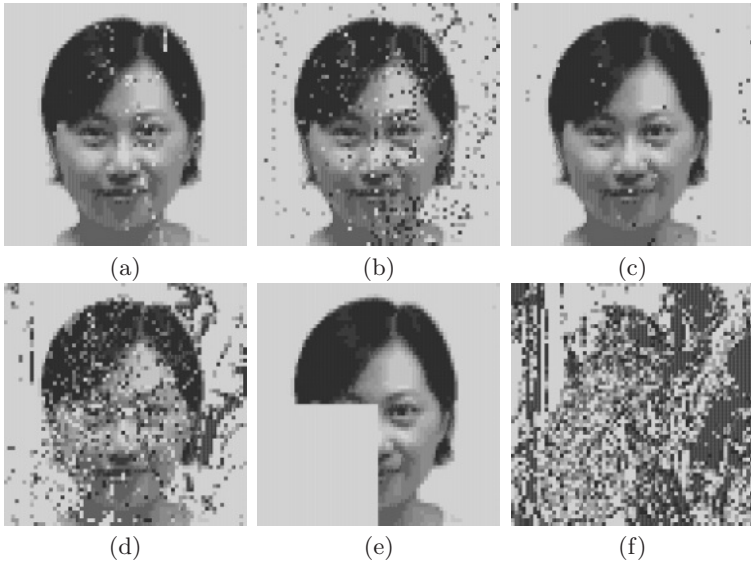


Fig. 5. The watermarks recovered from the attacked watermarked images while using LENA as the cover image and the GIA procedure introduced: (a) VQ compression with codebook size=512, (b) JPEG compression with QF=60%, (c) JPEG compression with QF=80%, (d) median filtering with window size=3, (e) 25% of cropping at the left-bottom corner, (f) using the original LENA as the input image for extraction

Table 3. Comparison of the related VQ-based watermarking methods in literature

Methods	Watermark size (bit)	PSNR (dB)	Note
VQ Coding	N/A	31.800	
[4]	128×128	28.245 (avg.)	Requires the original cover image during extraction.
[5]	$\leq 128 \times 128$ (16114.31, avg.)	28.653 (avg.)	A threshold of $D = 100$ was used.
[7]	128×128	31.800	Does not embed the watermark into the cover image.
Proposed	$256 \times 256 \times 8$	30.016 (avg.)	Proposed in Sect. 2.
Proposed with GIA	$256 \times 256 \times 8$	30.529	Proposed in Sect. 3.

mentioned (e.g., the original codebook) were used as other test data. For simplicity, only one watermark bit was embedded in each block of the cover image. The codebook partition keys used in [4,5] and the proposed scheme were generated randomly without employing any other training technique (but followed the rules proposed in the corresponding methods). The simulations were carried out for 1000 times and the results of them are listed in Table 3.

5 Discussion

Generally speaking, the proposed watermarking scheme provides better performance than some related VQ-based watermarking methods in literature. Table 1 shows that the GIA procedure helps improve the imperceptibility of the embedding results since the signal of the gray watermark is modified to suit the signal of the cover image. The recovered gray watermarks have better visual quality even the PSNR values under some attacks are not high, as shown in Fig. 5. Table 3 demonstrates that the proposed method has the best ability to hide a watermark with larger size. The issues regarding key reuse and key delivery are also considered by some users but they are beyond the focus of this paper.

6 Conclusion

We propose a new VQ-based watermarking scheme which expands the size of the used watermark and the genetic index assignment procedure which improves the performance of the proposed watermarking scheme. Experimental results and the discussion show that the proposed methods are effective, novel, and possesses many advantages than other VQ-based watermarking methods in literature.

References

1. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital Watermarking. Morgan Kaufmann (2000)
2. Katzenbeisser, S., Petitcolas F. (eds): Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Norwood (2000)
3. Gersho, A., Gray, R.M.: Vector Quantization and Signal Compression. Kluwer Academic Publisher, London England (1992)
4. Lu, Z.M., Sun, S.H.: Digital Image Watermarking Technique Based on Vector Quantisation. IEE Electronics Letters, Vol. 36. (2000) 303–305
5. Jo, M., Kim, H.: A Digital Image Watermarking Scheme Based on Vector Quantisation. IEICE Trans. Inf. & Syst., Vol. E85-D. (2002) 1054–1056
6. Wang, F.H., Jain, L.C., Pan, J.S.: A Novel VQ-Based Watermarking Scheme with Genetic Codebook Partition. Proceeding of the 3rd International Conference of Hybrid Intelligent Systems, The IOS Press, Melbourne (2003) 1003-1011
7. Huang, H.-C., Wang, F.H., Pan, J.S.: A VQ-based Robust Multi-watermarking Algorithm. IEICE Trans. Fundamentals, Vol. E85-A. (2002) 1719-1726
8. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Massachusetts (1992)

Advanced Paper Document in a Projection Display

Kwangjin Hong and Keechul Jung*

School of Media, College of Information Science, Soongsil University,
Seoul, South Korea

{hongmsz,kcjung}@ssu.ac.kr

Abstract. Though computing environments are developed very rapidly, people use paper documents as ever. In this paper, we propose the Advanced Paper Document (APD), which expands the function of paper documents, based on the Augmented Desk system. To incorporate physical paper documents with digital documents in a projection display system without being required to use additional sensors, we use a gesture recognition method. The APD has advantages of paper and digital documents, and provides a user with a natural and intuitive environment. As shown by experimental results, the proposed the APD is applicable to provide an interactive computing environment.

1 Introduction

Projection displays are kinds of displays, which produce images of large size and high resolution. Tangible interactions on projection displays can provide rich opportunities for augmented reality and ubiquitous computing. The DigitalDesk [2], the AugmentedDesk [3], the EnhancedDesk [4], and Tangible bits [5] have been developed and demonstrated that the tangible user interaction enables in their systems.

The first three systems use both virtual objects and physical objects, and provide users more intuitive interaction by allowing them to use their own hands for direct and fine manipulation of both physical and virtual objects. On the other hand, in the Tangible bits system, the capability of recognizing real objects on the desk is rather limited comparing with previous systems. The system depends on a specially designed tag attached onto a real object to determine which object a user is using.

In this paper, we propose an Advanced Paper Document (APD). The paper document is inexpensive, handy to carry, and good to read. Otherwise the digital document has more information than the paper. The APD that we propose has advantages of two kinds of documents those are the paper and the digital. It provides a user with an interactive environment to incorporate physical paper documents into digital documents in projection display system without using additional sensors or markers. To provide more convenient user interface, we implement the intuitive system using hands. For this work, we detect foreground

* Corresponding author.

objects using color calibration and stereo information and classify the shape of the hand using a k-NN method. Then the APD selects the document that the user wants to show, in the projection display using his/her hand which is detected at the previous step. When the user selects the document, APD shows the “*menu box*” that consists of “Information,” “Capture,” “Edit,” and “Scan,” and executes a command that user selects in the menu box also using the user’s hand.

This paper is organized as follows. Section 2 describes the Advanced Paper Document including a method of hands detection in section 2.1 and shapes recognition in section 2.2. Section 2.3 shows the documents retrieval method. Section 3 presents the experimental results, and some conclusions are given in section 4.

2 Advanced Paper Documents

In this paper, we propose the APD which does not use additional sensors but use only the user’s hand. The user can see overlaid online information of the document which is selected by him/herself. Therefore it provides more intuitive and natural environment.

In the APD, we use a hand like a pointing device. To detect the hand in projection display, we use a geometry and color calibration, and 3D information of the hand. And, the APD recognizes the shape of the hand using k-NN method. After the APD detects the shape and position of the user’s hand, the user can select a document using the hand. The selected document is segmented into paragraph using the X-Y recursive cut algorithm [12] and is compared with saved images using an edit distance method and pixel matching algorithm. At last, the APD select a matched document, and show the menu box for the selected document. And the user selects a menu using a hand, and is provided an interactive environment. Fig. 1 shows a flowchart of the APD.

2.1 Hand Detect

To detect the position of hands, we use result images those are made by the foreground objects detection method. This method uses an input image to the

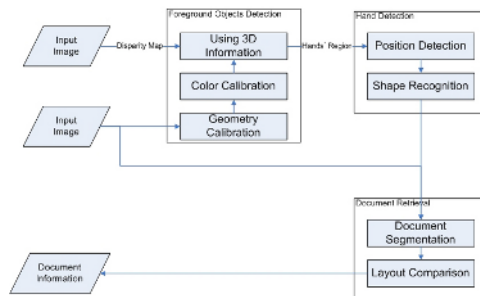


Fig. 1. A flow-chart of the APD.

projector, i.e. a *frame buffer image* and a captured image by a camera called a *camera image*. A camera image contains each instant scene on a projection display and foreground objects, with some distortions. Two major distortions are assumed: geometry and color distortion [1]. And we use 3D information of foreground objects from a stereo camera for the exactly detection.

To solve the geometry distortion, we use a projective transformation for coordinate transformation proposed by Ashdown M. et al. [6] and Sukthakar R. et al. [7]. Because it is reasonable to assume that both the projection display and corresponding camera image are in planar planes, the relation between the frame buffer image and the corresponding camera image is also assumed to be a projective transformation.

And, to solve the color distortion, we use brightness value variation. To detect brightness value variation, *brightness color markers* (BCMs) locate at the upper part of the projection display. BCMs have 10 blocks with 10-step gray values. Those are always shown in the upper part of the projection display.

To determine whether the pixel is a part of foreground objects or not, the color difference between the calibrated camera image and the frame buffer image is used. However, several errors appear in a certain environment. To acquire more robust results of the detection of foreground objects, we use the 3D information of foreground objects using a stereo camera. This 3D information is coded in the disparity map as brightness that bright pixels indicate near to stereo camera and dark pixels indicate away from the camera. Using the disparity map, we can improve the result of color-based foreground objects detection.

When a user selects the document in a projection display, the user uses two hands. However two hands in a projection display are not piled up. Therefore they can be separated using the Minimum Boundary Rectangle (MBR) of two hands each other. The nearest point from the left side of image, which is the result of the hands detection, and the point of the finger, which is the basis of the documents selection, is the same point. Therefore we can detect the position of the finger with the nearest point from the left side of image, and implement various commands.

2.2 Hand Shape Recognition

This system uses two kinds of the right hand shape to select a document or a menu. To classify the shape of hands in the region of right hand that is segmented at the previous step, we make features out of the run-length of black pixels through the row and column. And to recognize the shape of hands, we use a k-NN method. We get 20-features of the row and column direction respectively per one image as shown in Fig. 2.

And we classify two shapes using the clustering as shown in Fig. 3. In Fig. 3(a) and Fig. 3(b), graphes of upper side represent the variation of features of Fig. 2(a), and graphes of lower side represent those of Fig. 2(b). As shown in Fig. 2, we can easily detect the deference of the two classes in the feature space. Therefore, when the an image is inputted by a camera, we extract features of the hand, and compare features with the nearest value of each template. Then we can recognize a meaning of the shape.

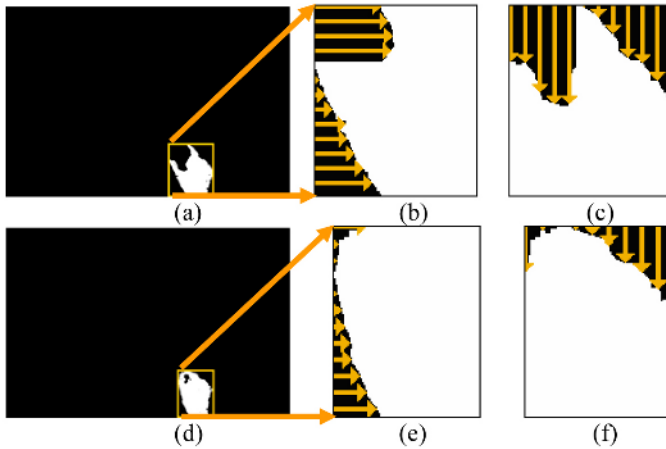


Fig. 2. Two kinds of hands' shape: (a) normal shape, (b) black pixel's run-length of the row direction of (a), (c) black pixel's run-length of the column direction of (a), (d) shape if the 'SELECT' operation, (e) black pixel's run-length of the row direction of (d), (f) black pixel's run-length of the column direction of (d).

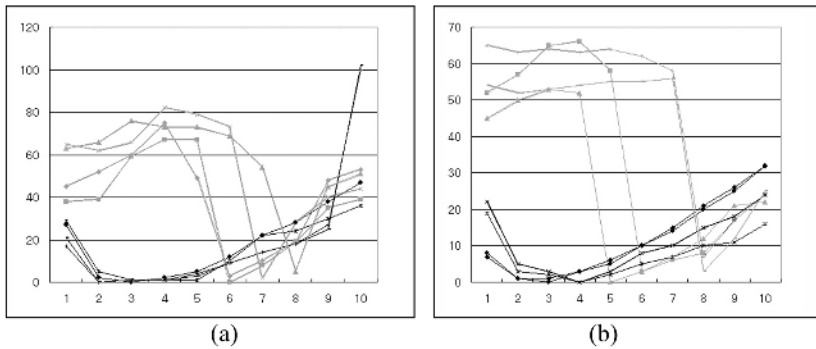


Fig. 3. Clustering of features: (a) black pixel's run-length of the column direction, (b) black pixel's run-length of the rowdirection.

2.3 Document Retrieval

For the document retrieval, previous studies have used string matching methods with using Optical Character Recognition (OCR) or structure similarity methods without using OCR. Also Studies of methods without using OCR can be divided into two methods such as using high-resolution images and low-resolution images. High-resolution images that can classify texts use a component blocks similarity method proposed by Peng H. et al. [10]. Otherwise, low-resolution images that can classify layout of the document use a layout similarity method proposed by Hu J. et al. [11].

Table 1. Average performance time of the document retrieval

	convert input image	compare with saved image	total
edit distance	171	396	567
pixel matching	168	572	740

In this paper, we use the low resolution binary image. Therefore we use the layout similarity method. It compares the size of the corresponding text region of the selected document in previous step with one of saved document images. And the document is decided the accepted document which has the highest similarity. To retrieve documents, we use the binary image of the camera image after calibrating distortions. Saved document images, which are compared with an input image, are scanned by a scanner and these are also binary images. To classify the text region in two images which are scan images and camera image, we use the X-Y recursive cut algorithm.

This system retrieves the saved document, which corresponds to the selected document, with comparing corresponded components. To calculate the similarity of documents, we use two methods and compare the performance and the speed between two methods. The result of comparison will explain in section 3.

The one is that the method using pixel matching algorithm compares corresponded pixels of the input image with those of saved images. The other is that the method using the edit distance, which proposed Hu J. et al. [11].

If there is the accepted document of the selected document, the menu box of the document is opened beside the paper document on the projection display by our proposed system. When a user selects a menu in the box, this system executes a command such as showing information about the document or the list of references of the document, and scanning the selected area of the document.

3 Experimental Results

The performance of the proposed method is evaluated in an experimental environment. The experiment system consists of a camera, stereo camera, a projector, a physical desk, and a standard PC. The image processing system consists of the CPU Intel Pentium4 2.66Hz, and the graphic card ATI Radeon 7000. The projection display is 1152×864 pixels with 32-bit true colors made by a projector BenQ HD2100. Camera images are captured by a Sony DCR-VX2000 camcorder and disparity maps are captured by a PointGrey Digiclops Vision System. The document is scanned by a HP Scanjet 5P scanner. The experimental system is implemented using Microsoft Visual C++ 6.0 and DirectX 9.0 SDK.

The APD captures scenes at speed of 5-frames per second, analyzes input images, and retrieves the selected image by user. As previously stated, we use two methods to retrieve documents. And as shown in Fig. 4, all of two methods show the satisfied result in document retrieval.

In the performance time, as shown in table 1, the method using the edit distance is faster than the method using pixel matching algorithm. Because we

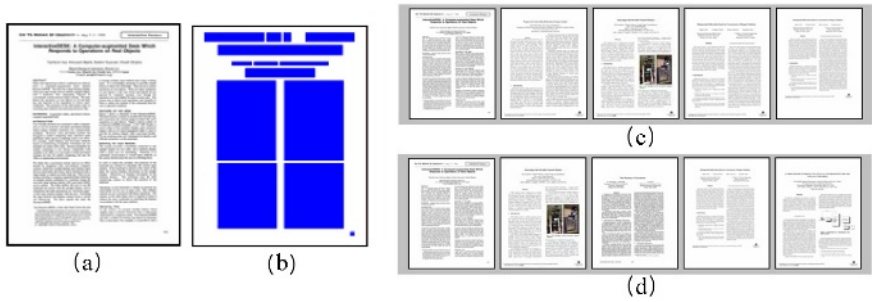


Fig. 4. The result of document retrieval: (a) input image, (b) segmented image of (a), (c) results of pixel matching, (d) results of edit distance.

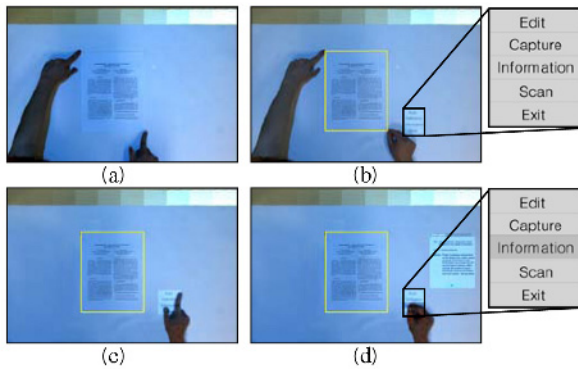


Fig. 5. Advanced Paper Document: (a) normal shape, (b) select the document and open the menu box, (c) select a menu, (d) click the menu button and execute the command of the menu (“Information”).

represents a document image as a 19×26 matrix, the number of comparing values are reduced. Therefore, because it shows ten-times decrease as compares with the method using the pixel matching algorithm, the method, which calculates the edit distance using dynamic programming, is faster than the method, which calculates the difference of pixels.

Fig. 5 shows steps that a user selects a document, and selects an “Information” menu, which shows the author, the title, and references of the selected document.

4 Conclusion

This paper proposed the APD. To provide a natural and intuitive environment to users, the APD used a hand likes a pointing device. For using the hand likes a pointing device, we need the point of the foreground object, and the

recognition of the hand's shape. And to add on-line information to a selected off-line document, we need a method for retrieving correct documents, which are identical with the selected document. Our proposed methods, which detect foreground objects, recognize the shape of hands, and retrieve documents, showed the satisfied results.

Otherwise, as previously stated, it is occurred delay when a user selects a document or a menu, since our proposed system captured scenes at speed of 5-frames per second. In further works, we will solve problems of the system's speed, and propose the method that provides more natural User Interface environment. And we will study the method that improves the resolution of the input image to revise the document and to provide a collaborative system.

Acknowledgement. This work was supported by the Soongsil University Research Fund.

References

1. Kang H., Kim S., Lee C., Jung K., Park M. H.: Foreground Object Detection in Projection Display. *Journal of The Institute of Electronics Engineers of Korea*, Vol. 41-CI, No. 1, (2004) 27-37
2. Wellner P.: The DigitalDesk Calculator: Tactile Manipulation on a Desktop Display. *Pro-ceedings of ACM Symposium on User Interface Software and Technology*, (1991) 27-33
3. Sato Y., Kobayashi Y., Koike H.: Fast Tracking of Hands and Fingertips in Infrared Images for AugmentedDesk Interface. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, (2000) 462-467
4. Koike H., Sato Y., Kobayashi Y.: Integrating Paper and Digital Information on Enhanced-Desk: a Method for Realtime Finger Tracking on AugmentedDesk System. *ACM Trans. On Computer-Human Interaction*, Vol. 8, Issue. 4, (2001), 307-322
5. Ishii H., Ullmer, B.: Tangible bits: towards seamless interface between people, bits and atoms. *Proceedings of CHI'97*, (1997) 234-241
6. Ashdown M., Robinson P.: The Escritoire: A Personal Projected Display. *Journal of WSCG*, Vol.11, No. 1, (2003) 33-40
7. Sukthankar R., Stockton R.G., Mullin M. D.: Smarter Presentations: Exploiting Homography in Camera-Projector Systems. *Proceedings on ICCV*, (2001) 247-253
8. Sato Y., Kobayashi Y., Koike H.: Interactive Object Registration and Recognition for Augmented Desk Interface. *Proceedings of ACM SIGCHI 2001*, (2001) 371-372
9. Robinson J., Robertson C.: The LivePaper system: augmenting paper on an enhanced table-top, *Journal of Computers & Graphics*, Vol. 25, Issue 5, (2001) 731-743
10. Peng H., Long F., Siu W.C., Chi Z., Feng D. D.: Document Image Matching based on Component Blocks, *ICIP*, Vol. 2, (2000) 601-604
11. Hu J., Kashi R., Wilfong G.: Document Image Layout Comparison and Classification, *15th International Conference on Document Analysis and Recognition*, (1999) 285-289
12. Jung K., Han J.: Hybrid approach to efficient text extraction in complex color images, *In Pattern Recognition Letters*, Volume 25, Issue 6, (2004) 679-699

Improving Web Browsing on Small Devices Based on Table Classification

Chong Wang^{2,*}, Xing Xie¹, Wenyan Wang², and Wei-Ying Ma¹

¹ Microsoft Research Asia, No.49, Zhichun Road, Beijing, 100080, P.R.China
{xingx, wyma}@microsoft.com

² Department of Automation, Tsinghua University, Beijing, 100084, P.R.China
wangchong99@mails.tsinghua.edu.cn
wyy-dau@tsinghua.edu.cn

Abstract. Browsing large web pages on small screens is still very inconvenient due to their limited display sizes. A straightforward solution is eliminating the annoying horizontal scrolling requirement, i.e. present all the contents into a single narrow column, usually named one-column view. It is implemented by deleting the layout that will cause horizontal scrolling. However, after deleting the layout structure, some structural information becomes unreadable, like data tables. In this paper, we propose an approach to improve web browsing on small devices by classifying HTML tables into data tables and layout tables. For data tables, we try to preserve the original structural information in the final presentation. Therefore, the browsing experience can be improved for the one-column view. Experimental results show that our approach can achieve a satisfactory performance.

1 Introduction

Recently, browsing webpages on mobile devices is becoming more and more popular. However, their usage for accessing the web today is still largely constrained by their small form factors. Webpage adaptation has received increasing attention. Many researchers and companies have provided various adaptation methods [2,3,4] to cope with the problem. Among them, the most famous one is one-column view browsing scheme, which has been implemented in several commercial browsers like Microsoft Pocket IE and Opera browser. In this scheme, the original layout is discarded if its width is possible to be longer than the screen width, while all the content is maintained. However, the method does not take one point into account - can the layout be discarded?

Webpages carry information in various ways - text, image, and video, etc. However some information may be correlated in a structured layout, such as stock quotes, shopping information and weather reports which are called tabular data or structured data. Figure 1(a) shows an example of tabular data. In HTML files,

* This work was performed when the first author was a visiting student at Microsoft Research Asia

correlated data often appear in table elements (data table). However, `<table>` tags are also frequently used to control the layout of the page (layout table), which means that the data are not strongly correlated. The detail definitions for these two types of table will be given in Section 2. Therefore, our task is to find the data tables, and try to preserve their original structural information during the transformation to the one-column view.

(a) A data table with columns: Symbol, Price, Change, Volume. Data rows include INTC, MSFT, CSCO, ORCL, and AMAT. A red circle highlights the 'Change' column.

(b) Displayed in the traditional one-column view. The data is stacked vertically. A red circle highlights the 'Change' column data.

Fig. 1. (a) A data table; (b) Displayed in the traditional one-column view

As we can see, table classification is critical for solving the problem. In [1, 7,8,9], researchers have proposed a few methods for classifying tables. However, there are still many problems. Their methods are not designed for the purpose of webpage adaptation. Some do not state a clear definition for various tables. And the algorithms proposed may be slow, and converge to local maxima [11].

In this paper, we present a table classification algorithm for one-column view based on machine learning, and the classification results are applied to improve the layout transformation from common webpages to the one-column view. Experimental results show that our method is much better than the traditional one-column view when webpages contain data tables.

The main contributions of this paper are:

1. A clear definition of data table and layout table for webpage adaptation is presented.
2. An improved one-column view approach is proposed to facilitate mobile web browsing.
3. A set of experiments have been carried out to verify our performance.

2 Data Table and Layout Table

We mainly focus on leaf tables [1] (A table element is said to be a leaf table if and only if it has no children table element in DOM tree hierarchy [7]), and our data tables are from leaf tables.

As mentioned in Section 1, the usage of table element in HTML file can be classified into two groups:

- Data table: the content in cells is clearly correlated. Figure 1(a) show an example of data tables. For this type of table, each cell’s relative position can’t be changed, or it will lose the meaning, since each cell can tell little without the cells in relative positions, as displayed in Figure 1(b). Say the number 26.470 in Figure 1(a), we know it is the price of INTC, while in Figure 1(b) we do not know what it is trying to say. The hierarchy is damaged and the validity of the information is lost.
- Layout table: these tables are only considered to be tables by virtue of their appearance once rendered by a suitable browser. In Figure 2, the circles in the same style indicate the corresponding areas. Clearly, we can read the content after transformation, since each cell doesn’t have much relationship with others. The functionality of a table element here is just to give a better appearance.



Fig. 2. (a) A layout table; (b) Displayed in the traditional one-column view

Sometimes, we may encounter tables in which the cells do have relationship, but this relationship is not strict. See Figure 3, the characteristic of this kind of table is that each cell plays an individual role; all the roles together form an entire meaning. Due to its individual role of each cell, we may split the table while maintaining the meaning. So we still call them layout tables.

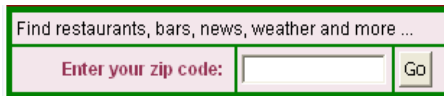


Fig. 3. A layout table that has a certain relationship among the elements

We cast the table classification as a traditional classification problem. The problem is to determine for each table element in the DOM a correct label: positive or negative. A positive label indicates that this table is a data table; otherwise it is a layout table.

Table 1. List of our features

Type	Feature name	Description
Visual features	Border width	The border width of the table
	Row span	Number of row spans
	Column span	Number of column spans
	Row and column bin features	Three bins representing the existence of 1-5, 6-10 and 11+ rows or columns
Content features	Textual content ratio	The ratio of cells with textual content to the total number of cells
	Singular cell ratio	The ratio of cells spanning exactly 1 row and 1 column to the total number of cells
	Link content ratio	The ratio of cells containing anchor texts to the total number of cells
	Image content ratio	The ratio of cells containing images to the total number of cells
	Digital content ratio	The ratio of cells with digits to the total number of cells

3 Table Features

Many features can be extracted to represent a table. In [1,7,8], the authors used heuristic methods to select a set of features that might be useful to the classification problem.

The features we use here are partitioned into two groups: visual features and content features. Our classifier works without rendering the HTML file. First we parse the HTML file into a DOM tree. Second, using the interfaces provided by the DOM parser, we get the features.

Table 1 illustrates the features in our experiments. Our features are an extended version of those introduced in [1]. Layout features used in [7] use the size and position information of table, which needs rendering of the HTML file; word group features require a large computational cost, which is not suitable for small handheld devices. Linguistic models typically describe natural language in terms of syntax and semantics. However using this kind of information in table classification is a non-trivial task. The situation becomes worse for mobile devices due to the limited computing power.

4 Experimental Results

Since the number of layout tables on the web is much larger than that of data table, we used keywords such as “stock”, “weather”, “report”, “shopping”, etc., searched them in Google, and checked the returned webpages, finally labeled the tables in the pages. We collected 314 tables with 191 layout tables and 123 data tables.

We ran a number of experiments to determine how different kinds of features benefited the training of classifiers. Various classification methods have

Table 2. Experiments using different classification methods

Methods	Features	Precision	Recall	F-value
Neural Network	V	82.3%	79.2%	80.8%
	C	92.2%	91.4%	91.8%
	CV	92.8%	91.8%	92.3%
SVM(linear kernel)	V	81.7%	78.3%	80.0%
	C	94.2%	92.4%	93.3%
	CV	94.6%	92.8%	93.7%
SVM(rbf kernel)	V	89.5%	87.2%	88.4%
	C	96.0%	94.6%	95.3%
	CV	96.8%	95.6%	96.2%

been proposed by the researchers in pattern recognition and machine learning areas. Among them we selected Neural Network [12] and Support Vector Machines [5,10] to do our experiments. We used five-folder cross validation, and the parameters of each classifier were well tuned.

The measuring criteria are Precision (P), Recall (R) and F-Value, which have been frequently used. In the training, we used either visual features (V), content features (C), or their combination (CV).

Results of our experiments are shown in Table 2 using different classification methods.

As we can see, the best results can be achieved by combining visual features and content features together and using support vector machines (rbf kernel). We finally trained the classifier for page adaptation using the whole data set.

Figure 4 shows a misclassified data table in our experiment: this data table is not well designed; there are too many blank cells.

Company Name (Symbol)	Price
SUN MICRO (JLIFE)	\$16.00 \$4.00
SUN MICRO (JLIFE)	\$12.00 \$12.00
SUN MICRO (JLIFE)	\$17.50 \$9.50

Fig. 4. A misclassified data table

Detecting structured data not formatted in `<table>` tags in webpages is also a non-trivial task. Because tags other than `<table>` usually do not provide a hierarchy structure to organize data, and it is difficult to generate features to represent the data. Luckily, structured data of this kind are not common, and we will leave it for our future work.

The time efficiency has also been tested in a simulated prototype on desktop computer, since the Pocket PC does not provide any programming interface of DOM. Our test bed is: CPU P4 2.8GHz, Memory 512M. The average time for

classifying a table is below 3.0 milliseconds using our final classifier. Therefore, we expect the speed will also be acceptable on mobile devices.

5 Page Adaptation Based on Table Classification

We have implemented a prototype of improved one-column view on desktop computers. After adding the module of the table classifier, we got our improved one-column view. If the table classifier specifies a table element as a data table, we will perform some special processing on it (will be discussed in detail afterwards); otherwise the table will be processed according to the original one-column rules. That is, if there is no data table detected, our improved one-column view is the same as the traditional one-column view. Figure 5 gives the work flow of our improved one-column view approach.

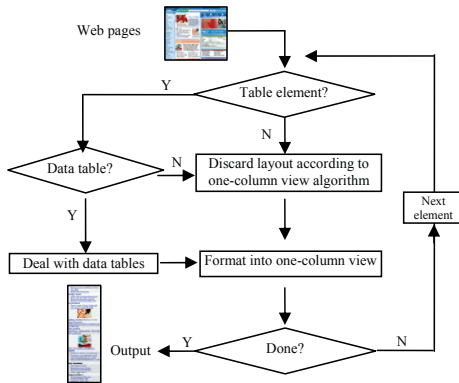


Fig. 5. Work flow of the improved one-column view

Dealing with data tables: As we know, the structures of data tables can't be discarded, so how to deal with data tables to prevent horizontal scrolling still remains a problem. Here we propose three approaches:

1. Do nothing, and keep it unchanged. Therefore, users may need to scroll horizontally when the table width is longer than the screen width.
2. Zoom the table in order to fit to the screen. Sometimes the content may become unreadable if the zooming ratio is too small.
3. Rotate the table, when its width is long, while its height is not. This transformation will not affect the structures, and it just changes the column wise [8] into row wise or vice versa. An example of column wise is the "Price" and the corresponding numbers in the same column in Figure 1(a).

Like doing a matrix transposing, the algorithm for rotating a table can be stated as follows:

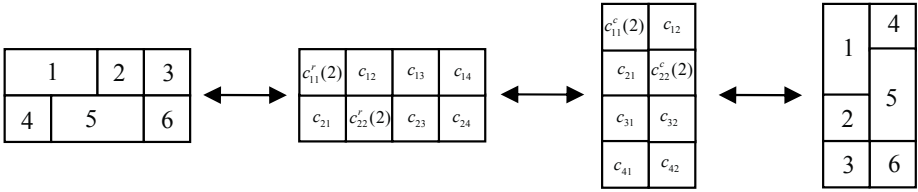


Fig. 6. Rotate a data table

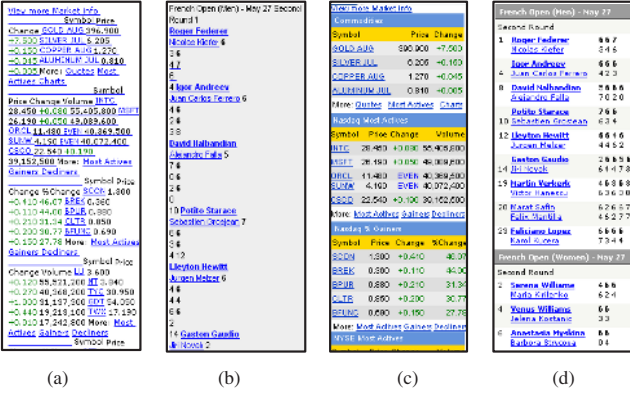


Fig. 7. (a) Traditional one-column view for the page at <http://www.stockhouse.com>; (b) Traditional one-column view for the page at <http://sports.yahoo.com>; (c) Improved result for the page in (a); (d) Improved result for the page in (b)

Assume a table has rows and columns, and let c_{ij} denote the cell in the i^{th} row and the j^{th} column ($1 \leq i \leq m, 1 \leq j \leq n$). If the table tags do not contain *COLSPAN* or *ROWSPAN*, the problem is easier. If there are tags with *COLSPAN* or *ROWSPAN*, we first drop *COLSPAN* and *ROWSPAN* by duplicating corresponding number of copies of cells in their proper positions. We introduce following symbols: $c_{jk}^r(k)$ denotes a cell in the i^{th} row and the j^{th} column which has *COLSPAN* = $k, k \geq 2$ and $c_{jk}^c(k)$ is defined analogously. In order to rotate a table, we just do the following mapping:

$$m \times n \rightarrow n \times m; c_{ij}^r(k) \rightarrow c_{ji}^c(k); c_{ij}^c(k) \rightarrow c_{ji}^r(k); c_{ij} \rightarrow c_{ji} \quad (1)$$

The new width and height will be recalculated to make sure that this transformation actually reduces the width of the original data table. A simple explanation of this kind of transformation is shown in Figure 6.

Some examples are shown in Figure 7. Clearly, our approach can provide a better presentation for these web pages on small devices.

6 Conclusion and Future Work

Common webpages are not suitable for displaying on small devices. One-column view is a simple transformation to facilitate the mobile web browsing. However, discarding most of layout information will cause the content elusive if the original data are correlated to each other in tabular structures.

We present an improved one-column view based on table classification. Our features include visual features and content features. These features are extracted without rendering the HTML files, which can be applied to traditional one-column view effectively without adding too much computation cost. The results show our method can achieve a satisfactory performance.

In the future work, we will consider how to detect those data tables not in `<table>` tags and how to show data tables in the one-column view more appropriately.

References

1. Hurst, M.: Classifying Table Elements in HTML. In: Proc. WWW'02, Honolulu, Hawaii, USA (2002)
2. Opera Small Screen Rendering.
<http://www.opera.com/products/smartphone/smallscreen/>
3. Chen, Y., Ma, W.Y., and Zhang, H.J.: Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices. In: Proc. WWW'03, Budapest, Hungary (2003)
4. Hori, M., Kondoh, G., Ono, K., Hirose, S., Singhal, S.: Annotation-based Web Content Transcoding. *Computer Networks* **33** (2000) 197–211
5. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines (2001). Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
6. Zanibbi, R., Blostein, D., Cordy, J.R.: A Survey of Table Recognition: Models, Observations, Transformations, and Inferences. *IJDAR* (2004)
7. Wang, Y.L., Hu, J.Y.: Machine Learning Based Approach for Table Detection on the Web. In: Proc. WWW'02, Honolulu, Hawaii, USA (2002)
8. Chen, H.H., Li, C.J., Tsai, J.H., Tsai, S.C.: Mining Tables from Large Scale HTML. In: Proc. ICCL'00, Saarbrücken, Germany (2000)
9. Yoshida, M., Torisawa, K., Tsujii, J.: A Method to Integrate Tables of the World Wide Web. In: Proc. WDA'01, Seattle, WA, USA (2001) 31–34
10. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (1995)
11. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **39** (1977) 1–38
12. Laurene V. F.: *Fundamentals of Neural Networks*. Prentice Hall, 1st Edition (1994)

A Java-Based Collaborative Authoring System for Multimedia Presentation

Mee Young Sung and Do Hyung Lee

Department of Computer Science & Engineering, University of Incheon, 177
Dohwadong, Namgu, 402-749 Incheon, South Korea
{mysung,oldkill}@incheon.ac.kr

Abstract. In this paper, we propose a Java-based collaborative authoring system for multimedia presentation. Our system is composed of an Editing System, a Media Object Manager and a Collaboration Manager. The Editing System contains two unique editors; a 3D Spatio-Temporal editor and a Temporal Relation Network (TRN) editor, in addition to the traditional editors such as timeline editor, tag editor, attribute editor, and text editor. These editors are all shared over the Internet and together they form an easy and efficient multimedia authoring environment. Using our system, users in different places can author multimedia presentations in a unified spatio-temporal space while freely traversing the spatial domain and the temporal domain without changing the context of authoring. We also implemented some ideas for efficient concurrency control in our system. They are mainly based on user awareness, multiple versions, and access permissions of shared objects. Our work on the development of a Java-based collaborative authoring system leads us to conclude that the Java technology is a satisfactory tool for developing a collaborative authoring system including a 3D interface. However, the Java3D technology needs to be ameliorated in the aspect of performance.

1 Introduction

Various editing facilities are needed for providing an efficient authoring environment. Among other things, a multimedia authoring system must provide an environment where temporal relations and spatial relations can be edited simultaneously. An interactive multimedia authoring system also needs to support user interactions. Some media (such as video, sound, and animation) require users to specify temporal characteristics, and other media (such as video, animation, images, and text) require users to specify the spatial relationship between objects.

The key to authoring a presentation lies in the composition of spatial relationships and temporal relationships between objects. The existing authoring tools usually provide an authoring environment where the spatial information and temporal information are edited independently in two different 2D GUIs (Graphical User Interfaces). One of the GUIs represents the spatial characteristics of the multimedia content and the other represents its temporal characteristics. Traversing two different GUIs can inconvenience the users.

Moreover, a 2D interface is not sufficient to completely represent multimedia information. Because, the spatial domain itself is two-dimensional, a 2D presentation space cannot accommodate the characteristics of both the spatial and temporal domains at the same time. Adding the temporal dimension to the two spatial dimensions results in three dimensions. We can represent simultaneously the 2D space and the 1D time in three-dimensional space. Therefore, a multimedia authoring tool benefits greatly from a 3D interface which intuitively represents the multimedia content in one seamless environment. Using a 3D interface, multimedia authors can display and edit the complete spatial and temporal information simultaneously.

We developed a collaborative multimedia authoring system based on the SMIL (Synchronized Multimedia Integration Language). SMIL is an XML-based markup language for integrated streaming media. The existing SMIL authoring tools provide basic user interfaces such as the scaled timeline-based user interfaces (representing media objects as different bars arranged in multiple layers on the scaled timeline) or textual tag editing user interfaces for authoring. What distinguishes our system from others is that it provides a unified 3D interface that allows for simultaneous authoring and manipulation of both the temporal and spatial aspects of a presentation.

In this paper, we propose a Java-based collaborative authoring system for multimedia presentation that is efficient through providing various editing facilities including concurrency control mechanisms. We will describe three major components of our system in the following section. The representation of temporal relations will be discussed in section 3. In section 4, we will examine the concurrency control mechanism. The implementation and the experiment of our system will be presented in Section 5 and 6 individually. Finally, the last section will provide conclusions and outline plans for future work.

2 System Structure

Our system represents images, videos, sounds, texts, text streams, and animations as 3D objects, and provides various editing functionalities for temporal compositions and spatial compositions. Our system is composed of the three main components: Editing System, Media Object Manager, and Collaboration Manager.

The Editing System consists of a 3D Spatio-Temporal Editor, a Temporal Relation Network (TRN) Editor, a Timeline Editor, a Tag Editor, an Attribute Editor, and a Text Editor. These editors exchange their information through the Media Object Manager and together form an easy and efficient editing environment. Among many editors, the 3D Spatio-Temporal Editor and the TRN Editor are special. The 3D Spatio-Temporal Editor represents media in 3D space. The Spatio-Temporal Editor is responsible for integrating and analyzing the detailed information about the spatio-temporal relationships among media. It allows users to edit the temporal relations between media via simple and direct graphical manipulations in 3D space in a drag and drop manner. TRN Editor vi-

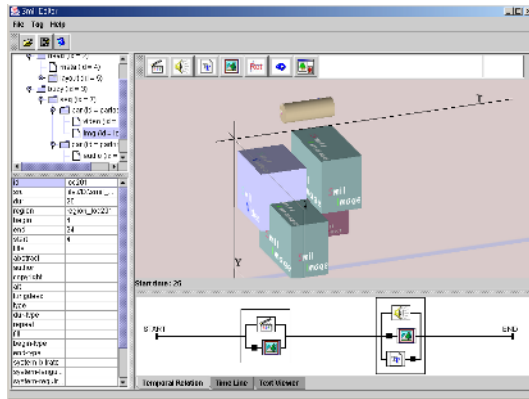


Fig. 1. An Example of 3D Representation of a Multimedia Presentation: Perspective View

visualizes the internal representation of the presentation and allows us to represent the conceptual flow of the presentation.

The Media Object Manager of our system is an essential part of our system. It is responsible for visualizing media objects in 3D, maintaining the consistency of the presentation information, and distributing all modification information from each editor to the other editors. The Media Object Manager consists of three modules: a 3D Engine, a Control Module, and a Parser & DOM Generator.

The Collaboration Manager allows a group of users working at different machines to work on the same multimedia presentation, and to communicate in real time. It processes the events and eliminates the collisions (the inconsistency of shared data which might occur when they are accessed concurrently). Our Collaboration Manager is composed of two modules: a Communication Module and a Concurrency Module.

3 Representation of Temporal Relations

Our 3D Spatio-Temporal Editor allows users in different places to simultaneously author and manipulate both the temporal and the spatial aspects of a presentation. Our authoring system represents a multimedia presentation in a 3D coordinate system. One axis represents the traditional timeline information (T-zone), and the other two axes represent spatial coordinates (XY-zone) as shown in Figure 1. Our system represents visual media object as 3D parallelepipeds and audio media objects as cylinders. The length of the shape along the time axis corresponds to the duration of the media. A cross section of the parallelepiped corresponds to the spatial area of the visual media to be presented.

Figure 1 illustrates a perspective view of a multimedia presentation. This presentation consists of five media objects, an audio clip, a video clip, two images, and a text object. Authors can create media objects, place objects in the desired

positions, and enlarge or shorten the temporal length of objects by dragging and dropping. A structural view of SMIL tags of the presentation (corresponds exactly to the DOM structure of the presentation) and a view of a SMIL object's attributes are also presented on the left of Figure 1. Authors can change the perspective from which the objects are viewed in 3D space using the arrow keys. Also, authors can quickly change to default views by selecting a corresponding icon. Details of our 3D Spatio-Temporal editor are described in the references [1].

Our system uses TRN (Temporal Relation Network) [1] as its internal representation of a multimedia presentation. TRN is based on Allen's temporal intervals [2]. Allen distinguished thirteen different time intervals between two objects. They can be reduced into seven temporal relationships such as 'before', 'meets', 'overlap', 'during', 'starts', 'finishes', and 'equals' by removing the relationships in inverse order. Our TRN supports these seven relationships. TRN corresponds exactly to the conceptual temporal structure of the multimedia presentation. TRN is a directed and weighted graph. Figure 1 also illustrates an example graphical representation of conceptual temporal relation of the presentation that is created as a user authors the presentation. After the authoring is finished, a DOM structure is generated from the graphical representation. The internal TRN and the graphical TRN are generated automatically from the 3D graphical representation specified by the author of the presentation. Our system finally generates SMIL documents through the interaction between the graphical representation and the DOM structure.

4 Concurrency Control

Collaborative systems need concurrency control to resolve conflicts between users' simultaneous operations [3]. There are many concurrency control mechanisms studied in the field of databases [4]. They are simple locking [5], transaction mechanism [6], turn-taking protocols [7], centralized controller [8], dependency detection [9], and reversible execution [10][11], operation transformation [12], multiple versions [13], etc. Finding a best concurrency control algorithm absolutely depends on the application semantics [14]. Also, it requires us to suffer from the tradeoff between the responsiveness and the performance for keeping the consistency of shared data. Our experience in the development of various collaborative systems leads us to conclude that a combination of the concurrency control methods can produce satisfactory results [1]. In our collaborative authoring system, we propose the following ideas for efficient concurrency control:

- Make users aware of every version of ongoing concurrent operations by changing the appearance of objects in concurrent access. One possibility is to give such objects a transparent look and show all concurrent operations. After the concurrent operations are complete, the proper version will be chosen as the final version is made visible to all users.
- Maximize the locking granularity by separating the temporal editing operations and the spatial editing operations of an object and applying different concurrency control mechanisms to each.

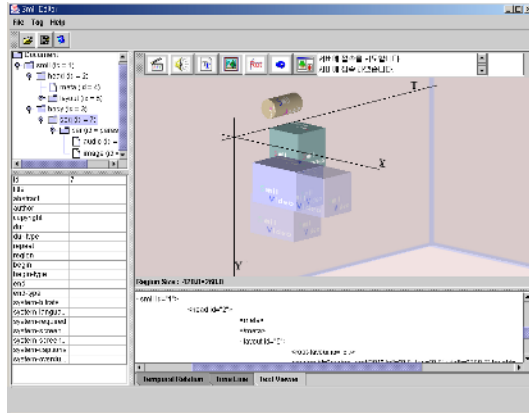


Fig. 2. An Example of Viewing Multiple Versions of Concurrent Editing

- Maximize the responsiveness using optimistic concurrency control with versioning, and minimize the collision due to the concurrent operations by requiring users to obtain access permission before editing.

Figure 2 presents an example of concurrent editing of an object. Details of concurrency control based on the access permission can be found in the reference [1].

5 Implementation

Our system is implemented using J2SDK (Java 2 Standard Development Kit) 1.4.1 and Java3D 1.3 library. The execution of our system requires an XML parser, since the grammar of the SMIL code is based on XML. Our system uses the JAXP (Java API for XML Processing) 1.1 for XML parsing. The collaboration module is implemented using Java sockets.

Figure 3 shows the structure of the main Java classes of our system. The role of each class is as follows:

- SmilEditor: This class manages and executes the overall system. It creates three classes for generating the interfaces. They are Inter3D, TreePane, and TablePane. The SmilEditor class also generates the classes for creating and manipulating the TRN (the internal representation of a multimedia presentation) and the classes for managing the network communications.
- Inter3D (means the 3D Interface): It generates the classes for managing the 3D Spatio-Temporal editor, such as SmilRoom. The class SmilRoom creates a 3D spatio-temporal space and also generates the RegionTrans class and MouseTrans class for the edition in the 3D spatio-temporal space. The RegionTrans class and the TimeTrans class take charge of transferring every editing result to the TRN structure. The Inter3D class generates the transparent classes and collaborates directly with the Network class in the case of collaborative authoring.

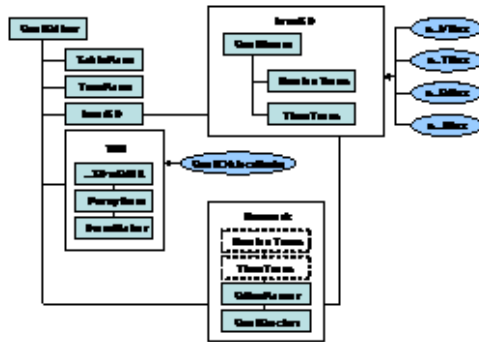


Fig. 3. Structure of Main Java Classes of Our System

- TreePane: It generates the classes for managing the Tag editor.
- TablePane: It generates the classes for managing the Attribute editor.
- TRN: This class creates a template of the SMIL document structure and modifies the values of the tag structure in real-time as the users author the presentation. It also generates a DOM structure and a SMIL file corresponding to the presentation using the `_3DtoSMIL` class, the `PrettyDOM` class, and the `DomMaker` class.
- Network: This class is generated when a collaborative authoring operation begins. It creates the shared media objects (transparently viewed) and works for the concurrency control in collaboration with the `Inter3D` class and the `TRN` class.

6 Experiment

We performed some experiments to validate the usability of our authoring system. The experiment was set with 10 university students who are good at computing. In the first experiment, the students proposed to author a multimedia content comprised of several media objects. We let five students to use the 3D editor (our system) first, then the 2D editor (TagFree 2000 SMIL Editor of Dasan Technology) second. We let another five students use the 2D editor first, then the 3D editor second. In the second experiment, we gave the students 6 different scenarios of presentations whose number of objects increases by 2 (such as 1, 3, 5, 7, 9, and 11) and let them author those presentations by crossing the editors in the same manner as the first experiment.

The results obtained through these experiments are as follows:

- The 3D authoring interface allows users to author faster than the two 2D authoring interfaces.
- The difference between the average authoring time for 2D and that of 3D increases as the complexity of media objects increases.

The first reason for latency comes from the time consuming adjustment of media objects on the scaled time-line for representing the temporal relationships

such as ‘parallel’ or ‘sequential’. The second reason originates from the change of the authoring environment from the temporal interface to the spatial interface, or vice-versa. The last reason is caused by the separate editing of layout regions in the spatial interface followed by the manual linkages of the layout objects to the objects in the temporal interface.

7 Conclusion

The objective of this study was to develop an easy and efficient collaborative authoring environment for multimedia presentation using Java technology. Toward this end, we created a shared authoring system which is composed of several editors such as a 3D Spatio-Temporal editor, a Temporal Relation Network (TRN) editor, a Timeline editor, a Tag editor, an Attribute editor, and a Text editor. Among these editors, the 3D Spatio-Temporal editor and the TRN editor are special.

In our 3D Spatio-Temporal editor, the spatial aspects and the temporal aspects of a multimedia presentation are represented in an integrated environment, so that users can create a multimedia presentation in a simple and intuitive manner. Our authoring system automatically converts the authored multimedia presentation to a Temporal Relation Network (TRN) for its internal representation. A TRN corresponds exactly to the conceptual temporal structure of the multimedia presentation. The internal TRN is visualized in the TRN editor. We also provide a concurrency control mechanism that is a combination of user awareness, multiple versions, and access permissions of shared objects. Our system allows users to compose and edit SMIL content in 3D space. In addition to their use in a SMIL authoring system, our system components can be used in a 3D virtual environment, for example a virtual collaborative system.

One advantage of our system is that multimedia authors need not change the editing environment from the spatial editing environment to the temporal editing environment, and vice-versa. Another advantage is its TRN representation of multimedia presentations. The TRN representation allows the system to automatically fill in the necessary timing details. This frees multimedia authors to focus instead on the creative aspects of the presentation. Also, we believe that the TRN representation can provide an efficient means for optimal automatic scheduling mechanisms to guarantee fine-grained synchronization.

We implemented our system using Java technology. Our work on the development of a Java-based collaborative multimedia authoring system leads us to conclude that the Java technology is a satisfactory tool for developing a collaborative multimedia authoring system. Java’s network capability is especially excellent. However, the Java3D technology needs to be ameliorated in terms of performance, even the programming with Java3D is easier than with OPENGL or DirectX directly.

In the future, we must further explore the concurrency control mechanism of shared objects to ensure their consistency and coherence. We want also to

examine the future extension of the SMIL standard for supporting authoring and the playback of 3D multimedia presentations in a 3D virtual environment.

Acknowledgement. This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Multimedia Research Center at the University of Incheon.

References

1. M.Y. Sung, D.H. Lee: A Collaborative Authoring System for Multimedia Presentation. Proceedings on The 2004 International Conference on Communications (ICC 2004) MM04-2 (2004) 1396-1400
2. James F., Allen: Maintaining Knowledge about Temporal Intervals. Communications of the ACM (November 1983) 832-843
3. C. A. Ellis, S. J. Gibbs, G. L. Rein: Groupware: Some Issues and Experiences. Communications of the ACM, **34**(1) (January 1991) 38-58
4. J. Gray, A. Reuter: Transaction Processing: Concepts and Techniques Morgan Kaufmann. (1993) 1070 pages
5. M. Stefik, D. G. Bobrow, G. Foster, S. Lanning, D. Tatar: WYSIWIS Revised: Early Experience with Multiuser Interfaces. ACM Transactions of office Information Systems **5**(2) (1987) 147-167
6. Dick C.A. Bulterman: SMIL 2.0 Part2: Examples and Comparisons. IEEE Multimedia **9**(1) (2002) 74-84
7. Mitsutoshi Iino, Young Francis Day, and Arif Ghafoor: An Object-Oriented Model for Spatio-Temporal Synchronization of Multimedia Information. IEEE International Conference on Multimedia Computing and Systems (1994) 110-119
8. B. Prabhakaran and S. V. Raghavan: Synchronization Models For Multimedia Presentation With User Participation. Proceedings on ACM Multimedia 93 (1993) 157-166
9. Naveed U. Qazi, Miae Woo, and Arif Ghafoor: A Synchronization and Communication Model for Distributed Multimedia Objects. Proceedings on ACM Multimedia 93 (1993) 147-155
10. Junewha Song, G. Ramalingam, Raymond Miller, Byoung-Kee Yi: Interactive authoring of multimedia documents in a constraint-based authoring system. Multimedia Systems **7** Springer-Verlag (1999) 424-437
11. S. Sarin, I. Grief: Computer-Based Real-Time Conferences. IEEE Computer **18**(10) (1985) 33-45
12. C. A. Ellis, S. J. Gibbs: Concurrency Control in Groupware Systems. Proceedings of ACM SIGMOD International Conference on Management of Data (1989) 399-407
13. C. Sun, D. Chen: Consistency Maintenance in Real-Time Collaborative Graphics Editing Systems. ACM Transactions on Computer-Human Interaction **9**(1) (2002) 1-41
14. J. Munson, P. Dewan: A Concurrency Control Framework for Collaborative Systems. Proceedings of ACM CSCW '96 (1996) 278-287

Object Tracking and Object Change Detection in Desktop Manipulation for Video-Based Interactive Manuals

Yosuke Tsubuku¹, Yuichi Nakamura², and Yuichi Ohta¹

¹ Department of Intelligent Technologies, University of Tsukuba, 1-1-1 Tennodai,
Tsukuba, 305-8573, Japan

tsubuku@image.esys.tsukuba.ac.jp

² ACCMS, Kyoto University, Sakyo, Kyoto, 605-8501, Japan
yuichi@media.kyoto-u.ac.jp

Abstract. This paper introduces a novel method for object tracking and recognition in assembly work. The purpose of this study is to index instructional videos and to provide appropriate instructions to a user during actual assembly work. Object tracking for this purpose involves a lack of prior knowledge such as an object's shape or color, since objects are often moved, assembled, or even crushed. The clutter present in an environment or environmental changes must also be addressed.

For this purpose, we use two or more pairs of image sensors. In this method, an object held by a hand is reliably detected, and its 3D area, that is, its volume and location, are obtained using shape-from-silhouette in real time. The observation of such volume allows the estimation of the changes in an object's state, and can be good indices for the processes of assembly work.

1 Introduction

The recognition of object manipulation such as that in assembly work has great importance in a variety of applications. For example, in assembly work, the detection of when and which object(s) is/are held, attached, or manipulated is essential in areas such as teaching, supervising, and recording.

As one of this type of recognition's promising applications, we are developing video-based interactive manuals for assembly work. If objects and important actions in videos are indexed and the current situation of a user can be recognized, we can provide a user with appropriate instructions by replaying the video portions. This will greatly help in the performance of the user's tasks and reduce accidents.

In this framework, however, difficult problems need to be addressed. It is difficult to provide the complete appearance of an object, since its appearance can be easily altered by rotation, deformation, or assembly; *e.g.*, joining parts. Moreover, it is natural that the object's background may change because of the inclusion of another person or moving object.

As one possible solution to this problem, we propose a novel method for detecting and tracking objects. In this method, we use two or more pairs of cameras, that is, an RGB camera and an IR camera. By using these cameras, we can effectively detect the part of the object which is held by hand even if prior knowledge of the object is not provided. The object's 3D position and its volume can be estimated by voxel carving. This information provides effective indices for an object and its changes, and increases the usability of the video manuals.

2 Object Recognition for Desktop Manipulation

2.1 Objectives

In this study we focus on cooking, assembly procedures, and scientific experiments. In such situations, an object's appearance or even its volume can often change due to rotation, the joining of parts, the separation of parts, and by partial occlusion.

Therefore, we consider object tracking under the following conditions:

- The system has no prior knowledge of an object's size, color, texture, etc.
- The background may change at any time during manipulation.

On the other hand, we can naturally make the following assumptions that will make these conditions easier to accommodate:

- Most of the important objects are moved or manipulated by human hands.
- The space (volume) in which important objects are likely to appear is known on the condition that the work space, *e.g.*, a workbench, is stationary.

Even with these assumptions, the conditions presented above are still severe. Although a number of studies have investigated hand tracking or object tracking, some of which have reported good results, our situation is much more difficult assumed in studies.

To cope with these problems, we earlier proposed a simple and robust method involving an RGB camera, an infrared (IR) camera, and a stereo camera[1,2]. This method enabled object detection and tracking in the above environment. However, there remain some problems to be solved, for example, the difficulty of recognizing the change/transformation of objects, the difficulty of tracking small objects, system cost, and so on. We, therefore, propose a new method for improving on these points.

2.2 Basic Idea

Our new method detects hands and objects, and also estimates the object's volume simultaneously. This volume estimation provides reliable detection of objects and their changes.

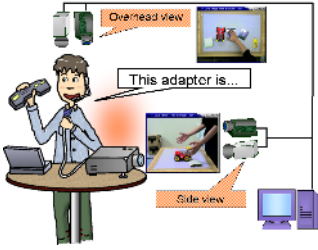


Fig. 1. System overview



Fig. 2. Images from RGB cameras and IR cameras

Figure 1 shows an example configuration of our system. Two types of cameras, that is, an IR camera and an RGB camera are located at each view point, *e.g.*, right above the work space, or on the left side of the work space. By detecting the “skin-color” region, the “skin-temperature” region, and the moving region, hand regions and held object regions are reliably detected. This portion is described in section 3. The actual 3D position of a held object regions are then detected in terms of voxel carving. After the volume is measured, we can estimate the changes in an object’s state by observing the changes in volume. This process is explained in section 4.

3 Detecting and Tracking of Hand-Held Object

3.1 Usage of Two Image Sensors

Figure 2 shows an example of images obtained by two types of cameras. The following types of regions are detected and utilized for detecting objects and recognizing their states.

hand region: A group of pixels that have both skin color in an RGB image and a pixel value corresponding to skin temperature in an IR image is considered as a hand region.

moving region / held object region: Frame subtraction and template matching are used for detecting moving regions. If a detected moving region is close to a hand region, it is a candidate for a held object region. The details of this process are explained below.

The rough idea governing the detection of those types of regions are denoted as follows:

$$\text{hand-region} = \text{skin-color-region} \wedge \text{skin-temperature-region} \quad (1)$$

$$\text{held-object-region} = \text{moving-region} \wedge \neg \text{hand-region} \quad (2)$$

3.2 Hand Region Detection

We created a skin-color model by gathering the statistics regarding pixel values showing skin color, and determined their distribution parameters. This method is based on the observation that skin color can be roughly modeled on the rg-plane¹. Although this approximation is not strict in that it accepts false alarms, we believe a strict model is not applicable here due to the presence of inter-reflections, shadowing, lighting changes, and so on.

First, the skin color regions are manually extracted, and the mean value and the covariance matrix Σ are calculated. Their actual values are as follows:

$$\begin{aligned} \text{mean}(\bar{r}, \bar{g}) &= [0.415483, 0.330955] \\ \Sigma &= \begin{bmatrix} 0.002025 & -0.000557 \\ -0.000557 & 0.000387 \end{bmatrix} \end{aligned}$$

The square of Mahalanobis distance $D^2(r, g)$ from skin color is calculated, and from this, the skin color regions are extracted.

$$D^2(r, g) = \begin{bmatrix} r - \bar{r} \\ g - \bar{g} \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} r - \bar{r} \\ g - \bar{g} \end{bmatrix}$$

The graph in Figure 3 shows the statistics obtained from a typical image in our environment. $D^2(r, g)$ values in actual skin regions and those in the background are plotted. Based on those statistics, we determined a threshold value of $Th_{\text{skin-c}} = 2.0$.

Next, our IR camera captures infrared light with a wavelength between 8 and $12\mu\text{m}$, which covers the dominant wavelength that the human body emits. We checked the pixel values in the real hand region and those in a typical background, and determined the threshold for extracting the skin temperature region.

In our experiments, a threshold $Th_{\text{skin-t}}$ of around 150 effectively separates those regions, as shown by the graph in Figure 4. Since the actual pixel value depends on the iris, focus, and other camera parameters, the threshold must be adjusted if those parameters are changed.

3.3 Held Object Region Detection

From the assumption mentioned in section 2.1, we consider a region moving in conjunction with a hand region as a held object. To detect moving regions, inter-frame subtraction is more reliable than is background subtraction, since we need to allow background changes during work.

In each frame subtraction image, the number of pixels each of which has a different value above the threshold th_d is counted for each window, *e.g.*, a 5×5 pixel window, placed throughout the image. If the number is larger than the threshold th_n , the window is considered a portion of the moving object. By gathering such windows close to a hand region, we can obtain the candidate

¹ A normalized color space. $r \equiv \frac{R}{R+G+B}$, $g \equiv \frac{G}{R+G+B}$.

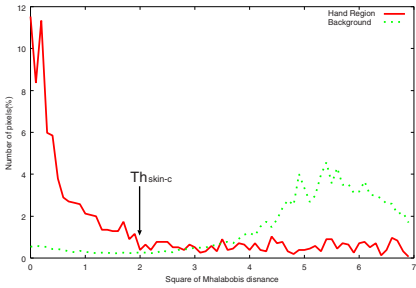


Fig. 3. Statistics of $D^2(r, g)$

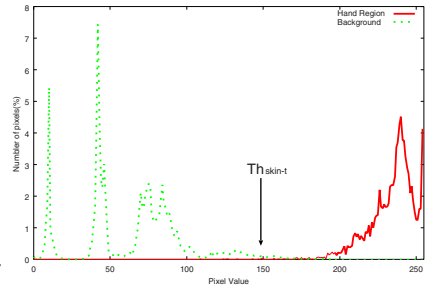


Fig. 4. Pixel values from an IR camera

regions of a held object. However, we cannot obtain a complete object region when an object stays still. In such a case, we use frame subtraction in conjunction with template matching as follows, since template matching is usually stable.

Step 1: If a held object region that is moving sufficiently fast is detected, the detected region is registered as the object template and this step is repeated. Otherwise, step 2 is performed. The speed can be checked by the position change of the object region. If the position change is larger than the object's size, the above condition is satisfied.

Step 2: The object's position is renewed by template matching with the template obtained by Step 1.

4 Detection of an Object's Change

For estimating the volume of a held object, we first need to delineate a held object region for each view point. For this purpose, since inter-frame subtraction is not sufficiently accurate for delineating an object's silhouette, we also use background subtraction. The background image is updated when moving regions are not detected in the frame subtraction image mentioned in the previous section. The volume corresponding to the held object can then be obtained by voxel carving[3,4].

In our experiments, $50 \times 50 \times 50$ (=125000) voxels each of which is $20 \text{ mm} \times 20 \text{ mm} \times 20 \text{ mm}$ in size are placed on the desk, and their computation is possible at 15 frame/sec on our PC (dual Xeon 3.06GHz).

Basically, the number of the voxels that correspond to the held object directly shows the object's volume. The volume can be used for recognizing the object, and can be also used for estimating its changes, *e.g.*, shape change, inflation, deflating/diminishing, and so on. It is, however, sensitive to noise, and it is difficult to discriminate whether the form of an object is actually changing or is merely affected by noise. For reducing noise effects, the system observes the volume for several frames, *e.g.*, in the sliding temporal window, then takes the second smallest value as the measured value. Then, by observing the volume changes, the system can also estimate the object's state changes.

In the case that someone is attaching two objects, there are no significant changes of the summation of the two objects' volume, and the two objects' regions move closer and finally join into one region. In contrast, both the summation of the volumes and the number of regions decrease when one object occludes the other.

We observe such changes and estimate which state change occurred in the objects. At its current developmental stage, this framework is not fully implemented, and we are now evaluating the capability of the volume measurements.

5 Experiments

5.1 Calibration

We need two types of calibrations for the cameras: one is the calibration between an RGB camera and an IR camera; the other is the calibration among view points.

Calibration between an RGB camera and an IR camera. The following quadratic transform $M_{3 \times 5}$ is used for considering 2D distortion and 2D perspective projection, where (x, y) represents the marker position in an RGB image, and (u, v) represents the marker position in an IR image.

$$\begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \\ 1 & \cdots & 1 \end{bmatrix} = M_{3 \times 5} \begin{bmatrix} u_1 & \cdots & u_n \\ u_1^2 & \cdots & u_n^2 \\ v_1 & \cdots & v_n \\ v_1^2 & \cdots & v_n^2 \\ 1 & \cdots & 1 \end{bmatrix}$$

We use 2D perspective projection, since the displacement between two cameras is small and the depth range of the workspace (desktop) is negligible in this context. Once correspondence is accurately obtained at some point in the workspace, it can be applicable to anywhere in the workspace.

Calibration among view points. By using the calibration pattern, we calculated perspective projection matrix $P_{3 \times 4}$ [5], where (u, v) represents the marker position in the RGB image or in the depth map, and (x, y, z) is the desktop coordinate system.

$$\lambda \begin{bmatrix} u_1 & \cdots & u_n \\ v_1 & \cdots & v_n \\ 1 & \cdots & 1 \end{bmatrix} = P_{3 \times 4} \begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \\ z_1 & \cdots & z_n \\ 1 & \cdots & 1 \end{bmatrix}$$

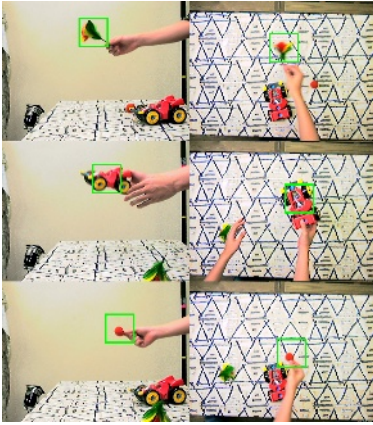


Fig. 5. Example of tracking result

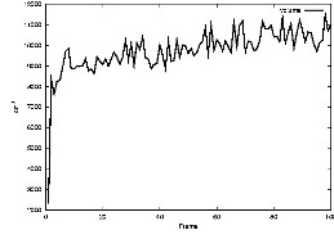


Fig. 6. Example of volume estimation (A cube whose volume is around 8000 cm^3 is slowly moved and rotated. Because of the inevitable drawbacks of the voxel carving, the estimated volume is slightly larger than the real value.)

Table 1. Operation detection performance

	flower	toy car	small ball
Exact	950 (95.0%)	920 (92.0%)	524 (52.4%)
Near	50 (5.0%)	54 (5.4%)	297 (29.7%)
Failure	0 (0.0%)	26 (2.6%)	179 (17.9%)
total	1000(100%)	1000(100%)	1000(100%)

5.2 Held Object Region Detection

Figure 5 shows the results of experiments regarding object detection and tracking in three cases: a toy car that is bigger than a hand, a flower that is a little smaller than a hand, and a small ball around 5 cm in diameter that is easily occluded by a hand. In those cases, a person held each of the objects and arbitrarily moved it.

Table 1 shows the accuracy of the tracking result, where “exact” means that the center of gravity of a detected region is on the object region, “near” means that a detected region is overlapping the object region, and the center of gravity is not in the object region. “Failure” means false alarm or false negative.

As we can see in the table, the detection rate (exact or near) even for a small object is above 80 %, and this rate is better for bigger objects. Considering that our system has no prior knowledge of the relevant objects, this is a very good result and would be difficult to achieve using a single image sensor.

5.3 Detection of Volume Change

For evaluating the accuracy of volume estimation, we performed two experiments: one regards measurement of an object whose volume is known; the other regards the measurement of a deflating object.

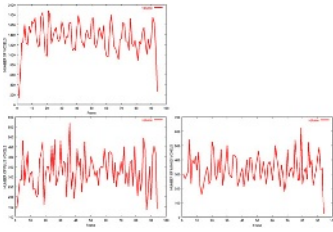


Fig. 7. Results of volume estimation by using background subtraction (The object was moved quickly. The upper left graph shows the estimated volume, the lower left shows the increase that is newly registered as a portion of the object, and the lower right shows the decrease that was transformed from the object to free space.)

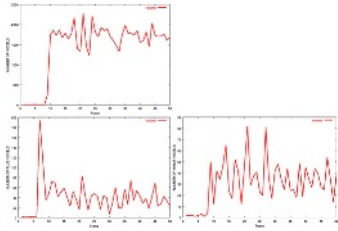


Fig. 8. Results of volume estimation result by using frame-subtraction (Compared to the case of background subtraction, increase and decrease are not accurately detected.)

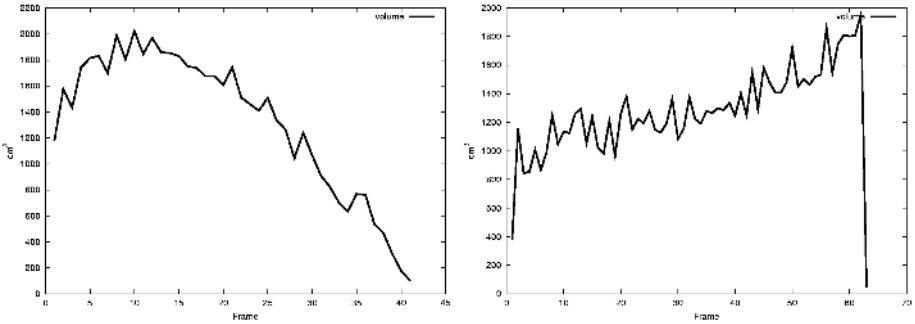


Fig. 9. Estimation of the volume of a balloon (The left graph shows the case of deflation, and the right graph shows the case of the balloon being punctured.)

For the first experiment, we used a cube whose sides were 20 cm, and moved it arbitrarily in the workspace. Since measurement is noise sensitive, we use the second largest value in five consecutive observations as the volume. Figures 6 and 7 show the results obtained by a simple method that uses background subtraction as shown in the previous section. The graph shows that the estimated volume is close to the correct value, though errors (over estimation) are not negligible. For comparison, the result of estimating volume by means of voxel carving with regions obtained by frame subtraction is shown in Figure 8. As we can see in the figure, the process of simply adding and subtracting the moving region is not reliable, and the above-mentioned method provides more accurate results. The details are omitted here due to lack of space.

The second experiment regards the measurement of changing objects. A balloon was held in the workspace, and it was gradually deflated or broken while it was moved by a hand. Figure 9 shows the results of volume estimation. The graph on the left-hand side shows the volume changes while the balloon was deflating, while the graph on the right-hand side shows the volume changes when the balloon was broken. Those graphs show that both phenomena were detected accurately, and we can easily distinguish between them.

These are still preliminary experiments, and we are currently attempting to improve the accuracy of volume estimation and more precise recognition of object changes.

6 Conclusion

In this paper, we proposed a novel method for detecting and tracking objects in desktop manipulation. We also described the principle of the estimation of an object's volume and recognition of an object's change of state. The preliminary experiments demonstrated the potential of our method. More advanced recognition will be realized in the near future. Though the system is still under development and needs improvement especially regarding accuracy, we believe that by increasing the number of viewing points this problem can be solved. The use of digital video camera with high resolution may also enhance the system's performance.

References

1. M.Ozeki, M.Ito, Y.Nakamura and Y.Ohta "Tracking Hands and Objects for an Intelligent Video Production System", *Proc. Int'l Conf. on Pattern Recognition*, pp.1011-1015, 2002.
2. M.Ito, M.Ozeki, Y.Nakamura and Y.Ohta "Simple and Robust Tracking of Hands and Objects for Video-based Multimedia Production," *IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems*, 2003.
3. S.M.Seitz and C.R.Dyer "Photorealistic Scene Reconstruction by Voxel Coloring," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1067-1073, 1998.
4. A. Laurentini "The Visual Hull Concept for Silhouette-Based Image Understanding," *Transactions on Pattern Analysis and Machine Intelligence*,16(2), pp.150-162, 1994.
5. J.Sato "Computer Vision -Geometry of Vision-," *Corona*, 1999.(in Japanese)

Design of an Integrated Wearable Multimedia Interface for In-Vehicle Telematics

Seongil Lee and Sang Hyuk Hong

School of Systems Management Engineering
Sungkyunkwan University
Suwon, 440-746, Korea
{silee, skkie96}@skku.ac.kr
<http://iesys.skku.ac.kr/~human/>

Abstract. A wearable, integrated wireless multimedia interface (WMI) for controlling in-vehicle telematics devices was developed in a Glove-typed interface using all the fingertips of both hands. The glove-based interface works for selection and control of devices using mode conversion of the respective keymaps. The chording gloves also work as a numeric keypad for mobile phones. To minimize finger force and fatigue from repeated finger tapping to the driving wheel, keys were made of conductible silicon ink. User evaluation on the WMI was positive with just a couple of hours of practice. The integrated wearable multimedia interface would play an important role in providing drivers with safety while driving.

1 Introduction

Multimedia systems have become important parts to automobiles. The increase in use of smart electronic products such as navigation systems, wireless internet, mobile phones, and other audio/video devices in automobiles provide convenient and pleasant driving conditions. Automobile infotainment systems for today's luxury cars provide noise-compensation and surround sound technology, communication systems with telematics and hands-free digital phone, navigation systems are equipped with DVD-based programs including maps and precise audio/visual directions. Control or manipulation of each device, however, requires human attention and inevitably distracts driver's attention from driving tasks, resulting in degrades in driving performance and potential danger of accidents. Emerging concept for central control of these multimedia devices suggests for needs of interface design that can provide straightforward, effortless, and logical activation and manipulation without driver's performance degrade in the main task of driving. This paper proposes an integrated and wireless wearable multimedia interface (WMI) in the form of chording gloves that can provide drivers with safe all-in-one telematics controller. Gloves equipped with a WPAN-based wireless communication module can be a good controller to multimedia systems inside automobiles since the hands and fingers are the most dominantly used parts of the body. Therefore, drivers can naturally develop unique and effortless strategies for interfacing with multimedia and telematics in the car without

moving their hands away from the driving wheel. In addition, the drivers do not have to move their eye gazes to control the devices while driving, which serves as a very important factor for safe driving.

1.1 Transport Telematics

A substantial amount of effort has been invested in transport telematics and in-vehicle multimedia systems over the last decade. Human interaction with computer systems has moved into the mobile environment and into everyday life, such as in-vehicle environment. Drivers can now get help to find the destination using in-vehicle GPS guidance systems, and check e-mail, appointments, or enjoy watching TV or listening to music using infotainment system. This change in use context already posed great design challenges and threw significant amount of design issues. The traditional display screen-keyboard interaction model is inappropriate for in-vehicle multimedia interface, and new interaction paradigms need to be investigated in multimodalities using speech or haptic input and output [1,2,9,13].

1.2 Chording Input Devices

Several types of glove-based devices recognizing hand gestures or contact gestures directly have been widely proposed as input devices to computers. These devices are well suited for use in a mobile environment because the gloves can be worn instead of just being used to hold a device, are lightweight and easy to store and carry. It is, however, difficult to recognize enough separate gestures to allow useful text input. Some glove-based input devices, though, have capabilities to make decent text input in addition to their intended functions of gesture recognition and space navigation. Pinch Gloves [3,12] are glove-based input devices designed for use in virtual environments, mainly for 3D navigation, and N-fingers [5] is a finger-based interaction technique for wearable computers also utilizing finger pinching. Pinching is basically the motion of making a contact between the tip of thumb and a fingertip of the same hand. It uses lightweight gloves with conductive cloth on each fingertip that sense when two or more fingers are touching. Pinch gloves were also used in a wearable computer for information presentation with an augmented reality user interface [12]. For use of many small portable electronic products, chord keyboards have also been proposed as input devices [3,4,5,8,10]. A chord keyboard is a keyboard that takes simultaneous multiple key pressings at a time to form a character in the same way that a chord is made on a piano. In chord keyboards, the user presses multiple key combinations, mainly two-letter combinations, to enter an input instead of using one key for each character. Pressing combinations of keys in this way is called chording [6,8,10]. Since chord keyboards require only a small number of keys, they do not need large space, nor the many keys of regular keyboards such as the QWERTY keyboard. For example, the Handkey Twiddler is a one-handed chord keyboard with only 12 keys for fingertips and a ring of control keys under the thumb, and the Microwriter with only 6 keys. Rosenberg and Slater [10]

proposed a glove-based chording input device called the chording glove to combine the portability of a contact glove with the benefits of a chord keyboard. In their chording glove, the keys of a chord keyboard were mounted on the fingers of a glove and the characters were associated with all the chords, following a keymap. Finge Ring [4] is another type of chord keyboard that uses finger movements with rings on the fingers instead of wearing gloves. Pratt [8] also designed a device-independent input device and language code called “thumbcode” using chording gloves targeting for PDAs. For extensive review on chord keyboards, see Noyes and other review papers [6,7,11]. All the previous works on glove-based input devices, however, lack consideration for accessibility and usability in mobile computing environments. None of the devices discussed above were targeted for use in multimedia systems, either.

2 System Structure

We developed an integrated and wireless wearable multimedia interface in the form of chording gloves that can provide drivers with safe, all-in-one telematics control functions. Gloves equipped with a WPAN-based wireless communication module can be a good controller to multimedia systems inside automobiles since the hands and fingers are the most dominantly used parts of the body. Therefore, drivers can naturally develop unique and effortless strategies for interfacing with multimedia and telematics in the car without moving their hands away from the driving wheel. In addition, the drivers do not have to move their eye gazes to control the devices while driving, which serves as a very important factor for safe driving. Control input can be made by simple tapping or chorded tapping of the fingers on the hard surface of driving wheel. The i-PAQ PDA with a Bluetooth module is currently working as a server for other in-vehicle information devices that are wired and as an access point for the wirelessly connected devices in our system (Fig. 1).

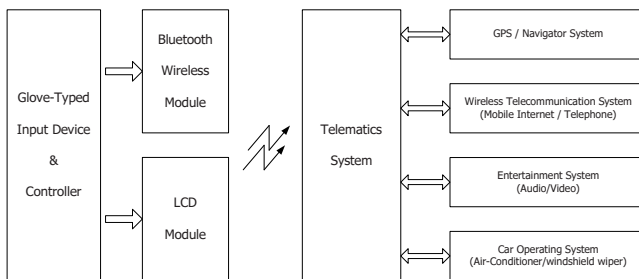


Fig. 1. System configuration for wireless access to in-vehicle telematics system

The design of glove-based WMI and its controller began with the arrangement of the sensors on the gloves which function as push-buttons. The interface of the

system is based on the human-friendly nature of finger movements, particularly finger tapping. All the operations by the chording gloves consist of either a series of finger tapping on a hard surface with a ground level.

2.1 Layout

A pair of chording gloves developed in the study which use finger tapping to make input is shown in Fig. 2. An input can be made by contacts between the keys on the fingertips and either the thumb or hard surface of the driving wheel. Chording is possible by making simultaneous tapping with multiple fingertips. Ten keys are placed on the fingertips of all the fingers including thumbs on the palm side of leather gloves. The keys are made of conductible silicon ink applied to the fingertips with the rectangle of 1.5 cm by 1.3 cm. The keys become “pressed” once the voltage through the silicon ink rises above 3.5 V with contact with any hard surfaces. The voltage outputs of chord gloves are connected to an embedded controller that translates chord information into its corresponding character. The corresponding finger force to generate 3.5 V is set to 3.50 N to avoid unwanted activations with light contact between the fingers and driving wheel. The chords are to be made as long as the two or more fingers tap simultaneously, and the system currently allows the time lag of 34.5 msec to be accepted as a simultaneous contact. The glove with the LCD panel and controller on weighs approximately 55.0 grams and the right-hand glove weighs 28.5 grams.



Fig. 2. The wearable multimedia interface (WMI) developed (left), and finger motions for making input using the WMI. Input can be made by simple tapping or by chording using two or more fingertips at the same time (center). Information of control mode and functions activated are shown on a small LCD of the left hand(right)

2.2 Controller

The 16F874 microprocessor is used for controlling and analyzing the signals from the WMI gloves. The controller also works for functions of converting the signals to text codes and sending them to a small LCD panel attached on the back of the left hand. The controller works with a 9V battery. The voltage outputs of chording gloves are sent to an embedded system that translates chord information into its corresponding code, number, or signals. The controller transmits

signals from the gloves to the PDA, mobile phones, and other telematics systems with a Bluetooth module to control them. The LCD panel presents text data transmitted from the gloves and status of the controlled device. It shows such information as status of selected device, controlling functions, and numbers dialed. Drivers need feedback from the system for what they activated and if they were correctly activated. In addition, two small vibrating motors attached on the back of the right-hand glove send a signal from the in-vehicle telematics devices to the driver for notices of device-initiated messages or for events to draw the driver's attention.

2.3 Keymap

The center of the wearable multimedia interface is its keymap with which control functions can be activated appropriately. The keymap for the WMI consists of a left-hand keymap and a right-hand keymap, where each hand conceptually plays a separate role.

Left-hand Keymap. The glove for the left-hand basically works as a device selection keypad. The index finger selects the mobile phone, while the middle finger selects the navigator on the iPAQ, and the ring finger for the A/V device. Once a device is selected, the interface changes to the control mode for the selected device until the other device is selected with the left-hand fingertip tapping. With the thumb excluded, a total of 10 devices can be selected from single finger tapping and two-finger chord tapping combinations. Pilot experiments showed that users could easily memorize and activate devices with chording combinations of adjacent fingers such as "index + middle" fingers or "ring + small" fingers, but had hard time to memorize and activate the functions assigned to the chords with non-adjacent fingers such as "index + ring" fingers.

Right-hand Keymap. The glove for the right-hand works for control functions for selected device and as a numeric keypad for the mobile phone. For example, once the CD player is selected by the left hand, then the index finger works for "play" and "pause", while the middle finger for "fast forward to the next track", the ring finger for "rewind to the previous track", and the small finger for "stop" functions. The two-finger chords are also used for volume control: "index + middle" finger chord for "volume up" control and "middle + ring" finger chord for "volume down." To maintain consistency, which is very important in designing a human-computer interface, the volume-related chords work for all the A/V device selections. Once a mobile phone was selected from the left hand, the right-hand fingers become numeric keypad, and the "thumb + index" finger chord works as a "send" function, and the "thumb + middle" finger as a "done" function. The information about the selected device and mode is displayed on the LCD on the back of the left hand glove. Right-hand keymap for selected devices are summarized in Table 1.

Table 1. Finger chordings and corresponding functions of the right-hand keypad

Finger chording	mobile phone	CD player	Navigator	Contact Error(%)
thumb	numeric 1	eject	guidance	1.25
index	numeric 2	play/pause	destination	1.25
middle	numeric 3	FF	starting position	3.13
ring	numeric 4	rewind	passing position	5.63
small	numeric 5	stop	exit/return	5.63
thumb+index	send		route search	5.00
thumb+middle	done		location search	7.50
thumb+ring	record		name search	10.63
thumb+small				18.75
index+middle	numeric 6	volume up	scale up(+)	2.50
index+ring	numeric 9			13.75
index+small	*/#			16.25
middle+ring	numeric 7	volume down	scale down(-)	7.50
middle+small	numeric 0			20.63
ring+small	numeric 8	repeat	new search	15.63

3 User Evaluation

A key to success for the chording gloves to be used in real-life in-vehicle environments as a multimedia control device would be its reliability and usability. To measure the reliability of the system, we measured the contact error rates for each finger contact and chorded contact. Eight college students (4 males and 4 females) with the average age of 23.3 old voluntarily participated in the experiment. The results are shown in the last column of Table 1. The numbers are the mean of error rates of the right and left hand fingers and chording. Users were also questioned on the WMI prototype's effectiveness, learnability, and potential distraction from safe driving. The ring and small fingers of both hands showed poor contact performances in both single finger contacts and chorded contacts in which they were involved. In chorded contacts, chording with adjacent fingers such as index and middle fingers showed much reliable tapping performances than ones with non-adjacent fingers such as a chording with index and ring fingers. The ring and small fingers apparently need longer time than the current setting for the contact, especially for chorded contacts, and the longer lag allowance is needed to make the system more acceptable and reliable. On average, it took approximately more than three hours of practice for the users to memorize the keymaps for both hands. Users found the experience of the WMI to be an interesting alternative to the traditional dashboard interfaces. Users found it extremely satisfying not to divert their eyes from the road to turn on the selected device and to control its functions. Users also rated the usability of the LCD panel high to see the information without diverting their eyes much, though many users also expressed a desire for larger font sizes or graphic symbols for effective feedback and communication. Most users said they enjoyed tapping



Fig. 3. A glove-based Wearable Multimedia Interface for wireless control of a navigation system on an iPAQ PDA and in-vehicle multimedia systems

the driving wheel to control the devices, but also pointed that wearing gloves on both hands is inconvenient and uncomfortable. Multimodal feedback using vibrations seemed to be effective to alert users while driving, though some users found it annoying.

4 Discussion

The wearable multimedia interface (WMI) in the form of chording gloves that were developed in our research for controlling in-vehicle telematics devices showed promises and problems. The WMI have distinct advantages over conventional push-button based on-board control panels of individual devices. The LCD of the current WMI provides only text data, which many users experience inconvenient. The gloves seemed to be a little heavy, especially the one with the LCD panel and controller. In addition, one-handed interface needs to be developed to free the other hand. The chording gloves will require extensive training with additional multimedia devices, and give users a load on working memory due to the absence of markings on the keys. The gloves also need to be tailored to each driver to minimize the discrepancy between the user's hand and the glove's size in order to minimize tapping errors. Above all, the WMI provides an all-in-one style control interface that does not distract driver's eye gaze from the normal line of sight required from driving. Since the interface can call for and control many in-vehicle multimedia systems without taking hands off the driving wheel, it provides drivers with safe driving environment. The LCD panel attached on the back of the left-hand glove mostly remained within 15 degrees of line-of-sight vertically while driving. The keymap for our chording gloves is quite simple and easy to learn, even though its interface needs to be investigated more in terms of human factors such as consistency and intuitiveness. The keymap can be altered at any time to satisfy any user's requirement. This would provide much flexibility in terms of usability. More detailed studies on interface design of the WMI must be followed to provide more driver-friendly in-vehicle multimedia interface system. However, the opportunity is sufficiently new for us to deal with

emerging in-vehicle telematics products and integrate them into a multimedia interface in the most effective ways we have.

Acknowledgements. This work was supported by grant No. R01-1999-000-00339-0 from the Basic Research Pro-gram of the Korea Science & Engineering Foundation.

References

1. Akesson, K. and Nilsson, A.: Designing Leisure Applications for the Mundane Car-Commute. *Personal and Ubiquitous Computing*, 6 (2002) 176-187.
2. Alpern, M. and Minardo, K.: Developing a Car Gesture Interface for Use as a Secondary Task. *Proceeding of CHI 2003 (Ft. Lauderdale, Florida, USA)*, ACM Press, (2003) 932-933.
3. Bowman, D. A., Wingrave, C. A., Campbell, J. M., Ly, V. Q.: Using Pinch Gloves for both Natural and Abstract Interaction Techniques in Virtual Environments. In: *Human-Computer Interaction; Proceedings of the HCI International 2001*, Lawrence Erlbaum Associates, (2001) 629-633.
4. Fukumoto, M., Tonomura, Y.: "Body Coupled Finge Ring": Wireless Wearable Keyboard. *Proceedings of CHI '97 (Atlanta GA, March 1997)*, ACM Press, (1997) 22-27.
5. Lehikoinen, J., Roykkee, M.; N-fingers: a finger-based interaction technique for wearable computers. *Interacting with Computers*, 13 (2001) 601-625.
6. Noyes, J.: Chord keyboards. *Applied Ergonomics*, 14(1) (1983) 55-59.
7. Porosnak, K. M.: Keys and Keyboards. In: Helander, M. (ed.): *Handbook of Human-Computer Interaction*, Elsevier, New York (1988) 475-494.
8. Pratt, V.: Thumbcode: A Device-Independent Digital Sign Language. *Proceedings of the 13th Annual IEEE Symposium on Logic in Computer Science*, Brunswick, NJ (1998).
9. Pieraccini, R. et. al.: Multimodal Conversational Systems for Automobiles. *Communications of the ACM*, 47(1) (2004) 47-49.
10. Rosenberg, R., Slater, M.: The Chording Glove: A Glove-Based Text Input Device. *IEEE Trans. On Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 29(2) (1999) 186-191.
11. Sturman, D., Zeltzer, D.: A Survey of Glove-based Input. *IEEE Computer Graphics & Applications*, Jan (1994) 30-39.
12. Thomas, H., Piekarski, W.: Glove Based User Interaction Techniques for Augmented Reality in an Outdoor Environment. *Virtual Reality*, 6 (2002) 167-180.
13. Wheatley, D.: Beyond the Desktop - and Into Your Vehicle. *Proceedings of the CHI 2000*, ACM Press (2000) 43-44.

Video Scene Retrieval with Sign Sequence Matching Based on Audio Features

Keisuke Morisawa, Naoko Nitta, and Noboru Babaguchi

Graduate School of Engineering, Osaka University
2-1 Yamadaoka Suita, 565-0871 Japan
{morisawa,naoko,babaguchi}@nanase.comm.eng.osaka-u.ac.jp

Abstract. In this paper, we propose a method of quickly retrieving semantically similar scenes to a query video segment from large-scale videos with audio features. This method first classifies the sound of the target and query videos into voices and background sounds and extracts feature vectors by focusing on the sound sources. The feature vectors are then clustered by K-means algorithm and the cluster ID, which we call *sign*, is assigned to the feature vectors in the corresponding cluster, consequently representing a video segment as a *sign sequence*. Finally, the video scenes are retrieved by sign sequences matching using Dynamic Programming. The experimental results show this method is potentially useful for scene retrieval.

Keywords: Sign sequence, audio feature, scene retrieval.

1 Introduction

Recent development of the information communication technology raises the need for an appropriate and flexible handling technique for large-scale video data. Content-based video scene retrieval[1][2] is an example of these techniques. Nevertheless, in order to retrieve the scenes which we want to see from the video stream under existing circumstances, we can only search them by time-consuming viewing. One of the techniques for solving this problem is to give the indexes which reflect the semantic content to video segments. Several indexing methods with the visual and textual features have been proposed until today [3][4], however, now the audio signal is also considered to play an important role in content analysis of audiovisual data[5] since the audio is closely related with the semantic content of the video.

Some methods for video indexing with the audio information have been proposed. Liu *et al.* [6] applied audio analysis results to make the distinction among five different video scenes: news reports, weather reports, basketball games, football games and advertisements. T. Zhang *et al.* [5] classified audio signals into speech, music, and other classes based on energy, zero crossing rate, and fundamental frequency features. Patel and Sethi [7] proposed to perform audio characterization on MPEG compressed data for the purpose of video indexing. The audio signals were classified into dialog, nondialog and silence intervals. An

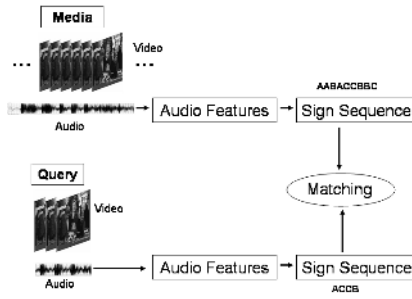


Fig. 1. Outline of the proposed method

approach to index videos through music and speech detection was proposed by Minami *et al.*[8]. Though all of these methods assigned the sound sources as video indexes, it takes time to specify the sound sources. Therefore, we propose an easier and faster indexing method by making more symbolic distinction among the audio segments.

Our proposed method is designed as follows. First, we classify the fixed-length audio segments into two classes: voices and background sounds, and then extract the audio feature vectors which characterize the sound sources of the background sound per one second. The audio feature vectors are classified into a specific number of clusters and are assigned cluster IDs, which we call *signs*. As a result, the audio sequences are transformed into the sign sequences as video indexes and the video scenes which have the similar audio streams to the query video segment are retrieved by matching the query and the target sign sequences.

The rest of this paper is organized as follows. Section 2 outlines the proposed method. Section 3 presents audio features and sign sequence matching. In Section 4, we show experimental results. Section 5 gives concluding remarks.

2 Outline of the Method

Fig. 1 shows the outline of the proposed method. For voice analysis, a time window of msec. width called a *frame* is usually used. Therefore, we also analyze the audio stream per frame and classify frames into voice frames and background sound frames with their features. Voice frames contain human voices and background sound frames do not. The voice features are considered predominant in the frame where the voice overlaps with the background sound. Therefore, we regard the frame which contains both voices and background sounds as a voice frame. For sound sources recognition, we use one second window [9], called *packets*, as processing units and extract the features per one second.

Let us here consider the sound sources of the background sound. There are four categories for the background sound sources: cheer, music, noise, and silence. Paying attention to these sound sources, we extract the audio features which indicate the differences of the sound sources.

Next, clustering is performed with the sample audio features on the vector space and a sign is assigned to each generated cluster. We then transform the target and the query features into sign sequences and retrieve similar scenes by matching these sign sequences.

3 Scene Retrieval

To retrieve similar scenes to a query video segment, we need the following operations: extraction of audio features, transformation of features into sign sequences, and sign sequence matching. In this section, we describe the details of these operations.

3.1 Audio Features

We separate the frames into voice frames and background sound frames with the frame features to reduce the influence of voices on the sound sources. The first characteristic of voices is the larger amplitude than other sound sources. Therefore, we use short time energy (*STE*) which represents the amplitude. The second characteristic of voices is the large low frequency power called formant frequency. Since the sound of video media also has the characteristic that the amplitude of 440-4000Hz is large, we perform a short time spectrum (*STS*) analysis to extract the power of 440-4000Hz. This technique carries out Fourier transform not in the whole sound stream but in 10-20 msec. intervals shifting every 5-10 msec. repeatedly. The shift size is called a frame cycle F_c . We calculate the short time energy *STE* for each frame and the average value of the amplitudes in 440-4000Hz is calculated with the short time spectrum *STS*. Let L denote the total number of signals in a frame. $STE(n)$ and $STS(k)$ are defined as

$$STE(n) = \sqrt{\frac{1}{L} \sum_m [x(m)w(m-n)]^2} \quad (1)$$

$$STS(k) = \frac{1}{2\pi L} \left| \sum_{m=0}^{L-1} x(m)e^{-j\frac{2\pi}{L}km} \right| \quad (2)$$

where $x(m)$ is the discrete audio signal, $w(m)$ takes 1 when m is included in the time window and 0 otherwise, $STS(k)$ is the short time spectrum when the frequency is $\frac{kf}{L}$ ($k = 0, \dots, L-1$), and f is the discrete sampling frequency.

The frame which satisfies the following two conditions is classified as a voice frame and otherwise as a background sound frame.

- The value of the *STE* exceeds a threshold $Th1$.
- The amplitude of the *STS* from 440 Hz to 4000 Hz exceeds a threshold $Th2$.

The features of a packet are extracted in consideration of the sound sources of the background sound. Noise has the continuous low amplitude and the large zero crossing rate. Music has several following characteristics. The zero crossing rate is small, the spectrum characteristic tends to remain the same in some intervals,

and the short time spectrum STS is strong between 4 kHz and 11 kHz. Below are the audio features for the audio packets.

a) *Average short time energy, \overline{STE}* : The average of $STE(n)$ in a packet.

$$\overline{STE} = \frac{1}{N} \sum_{n=0}^{N-1} STE(n) \quad (3)$$

where N denotes the number of frames in a packet.

b) *Low STE ratio, $LSTER$* :

$$LSTER = \frac{1}{2N_B} \sum_{n=0}^{N_B-1} |sgn[\overline{STE} - STE(n)] + 1| \quad (4)$$

where N_B denotes the number of background sound frames in a packet.

c) *Average zero crossing rate, \overline{ZCR}* : The zero crossing rate $ZCR(n)$ is defined as

$$ZCR(n) = \frac{1}{2} \sum_m |sgn[x(m)] - sgn[x(m-1)]| w(m - nF_c) \quad (5)$$

where $sgn[x(m)] = 1$ if $x(m) \geq 0$; $sgn[x(m)] = -1$ otherwise. \overline{ZCR} is the average of ZCR for background frames.

d) *Spectrum flux, SF* :

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} |\log(STS(n, k)) - \log(STS(n-1, k))| \quad (6)$$

where $STS(n, k)$, $k = 1, \dots, K$ is the k th spectrum at time n .

e) *Average spectrum envelope power, $\overline{S_{env}}$* : Let S_{env} denote the power value obtained from the spectral envelope from 4kHz to 11kHz. $\overline{S_{env}}$ is the average of S_{env} for background frames. We define $\overline{S_{env}}$ as follows.

$$\overline{S_{env}} = \frac{1}{N_B} \sum_{n=0}^{N_B-1} S_{env}(n) \quad (7)$$

f) *Voice frame ratio, VFR* : The voice frame is determined as the frame whose STE and amplitude of the spectrum at low frequency are large. VFR is the ratio of the voice frames to all the frames.

For the audio packet, we form a six-dimensional feature vector as $(\overline{STE}, LSTER, \overline{ZCR}, SF, \overline{S_{env}}, VFR)$.

3.2 Sign Sequence Matching

We transform the sequences of audio packets into sign sequences. First, as stated earlier, the audio feature vectors are extracted from the sample videos from various genres, such as sports, news, movies, animations, and dramas. For the audio vectors in the feature space, K-means clustering algorithm, which produce K_{max} clusters, is operated. This method initially takes K_{max} components that are equal to the final required number of clusters. In this step, K_{max} components are chosen such that they are mutually farthest apart. Next, each component is assigned to the cluster with the minimum distance. The centroid is recalculated every time a component is added to the cluster and this continues until all

the components are grouped into K_{max} clusters. Signs A_k , $k = 1, \dots, K_{max}$ are assigned to the clusters in the order of distance from the origin of the feature space to the centroids of clusters. Since the performance of K-means algorithm depends on the initial centroids, it is not regarded as the optimal clustering algorithm. However, due to its simplicity and practicality, the algorithm has been the main technique for clustering.

Next, we consider generating the sign sequences. The target video from which we intend to retrieve similar scenes is represented as a sequence of packets. We extract features from each packet generating its feature vector. The closest cluster is found with the nearest neighbor method. Based on the Euclidean distance between the centroids of the clusters and the vector, the sign of the nearest cluster is assigned to the packet. The sign sequences for the target and the query videos are called '*target sign sequence*' and '*query sign sequence*' respectively.

The similar scenes to the query video segment can be obtained with sign sequence matching. However, in many cases, the length of the scene which we want to see in the target video and the length of the query video segment is not the same. Therefore, we perform the sequence matching based on dynamic programming in order to handle the difference in the sequence length. First, we produce a score matrix whose elements depend on the distance among the centroids of all clusters. Using this matrix, we determine the temporal position in the sign sequence where the edit distance is minimized. The *edit distance* is defined as the sum of the scores required for unification of the two sign sequences by repeating following two operations:

Insert: A blank is inserted in order to extend a sign sequence.

Exchange: The sign in one sequence is replaced with the sign in the other sequence.

This matching algorithm is frequently exploited in DNA sequence alignment. Finally, we can obtain the temporal position where the target scene begins.

4 Experimental Results

Experiments were conducted with actual broadcast videos whose genres were the Japanese entertainment and news programs to check if this method can be applied to various genres of videos. The parameters were set as follows. Audio signals were sampled at 48 Kbps and quantized by 16-bit, L is set to 512 (about 11 msec.) and F_c is set to 128 (about 3 msec.). The threshold which distinguishes frames between voice and background sound frames were experimentally determined as 1800 and 10 respectively i.e. $Th1 = 1800$, $Th2 = 10$. We accept less than 60 seconds difference between the start position of a retrieval result and that of a correct answer. We evaluated the clustering results allowing for a margin of error within 10, 30 and 60 seconds with two, four, eight, sixteen signs, i.e. $K_{max} = 2, 4, 8, 16$. Retrieval results were ordered according to their matching scores. Video segments which are located within time interval of the actual start position of the scene are considered to be detected correctly. We checked if there was any correctly detected video segment in the first three retrieved results. The

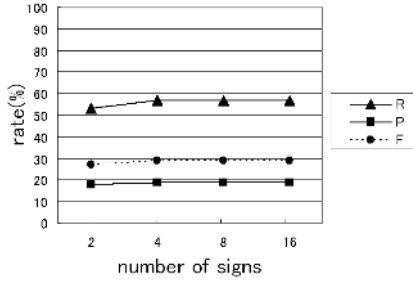


Fig. 2. Experiment A: R , P , and F within 10 sec.

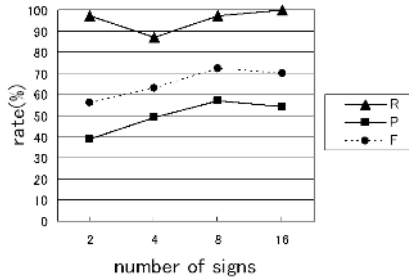


Fig. 3. Experiment A: R , P , and F within 30 sec.

measures for evaluation are the recall rate $R = \frac{\text{number of correctly retrieved scenes}}{\text{number of scenes to be retrieved}}$, the precision rate $P = \frac{\text{number of correctly retrieved results}}{\text{number of retrived results}}$, and the F-measure $F = 2 \cdot R \cdot P / (R + P)$.

Experiment A: Scene Retrieval from an Entertainment Program

We tried to retrieve similar scenes from a target video giving a query video scene from the same program broadcasted in a different day. The targets are eight 60-minute videos of a daily TV program, which were broadcasted in eight different days. We tried 32 retrievals in total with four queries. Each of eight target videos has a single scene that should be retrieved. We used the approximately 15-minute talk scenes as queries. Figs. 2 - 5 show the retrieval results.

Experiment B: Scene Retrieval from a News Program

We tried 30 retrievals in total with five queries and six target videos. The length of each target video was 60 minutes. We used the approximately 3-minute weather forecasts as queries. Figs. 6 - 9 show the retrieval results.

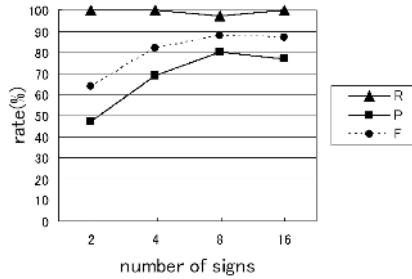


Fig. 4. Experiment A: R , P , and F within 60 sec.

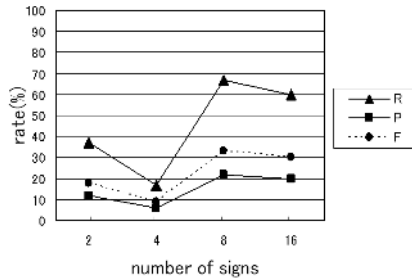


Fig. 5. Experiment B: R , P , and F within 10 sec.

Good results were obtained for scenes including music because the audio features used in this method were suitable for identifying music. However, the unstable results were obtained when the number of signs was two. This is because two signs are too few to represent the characteristics of the scenes, and too many similar segments to the query are extracted from the target video. For both experiments, when margin of errors was within 30 or 60 seconds and the number of signs was more than eight, we obtained relatively stabilized F-measure results. Even though the target video includes the same segment from the same TV program as the query video scene, these segments are not exactly the same and have different people, speakers, etc. The experimental results show that our method is flexible enough to successfully cope with such a rough approximation, which is considered to be difficult for exact matching methods such as speaker/sound source matching. Although the experiments we have conducted are rather simple and preliminary, we were successfully able to retrieve semantically similar scenes to the query at high speed, taking less than 1 second for sign sequence matching. As a future work, more evaluation using videos of different genres with more complex queries of different length is necessary to prove the effectiveness of our method.

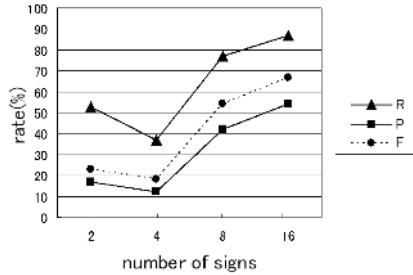


Fig. 6. Experiment B: R , P , and F within 30 sec.

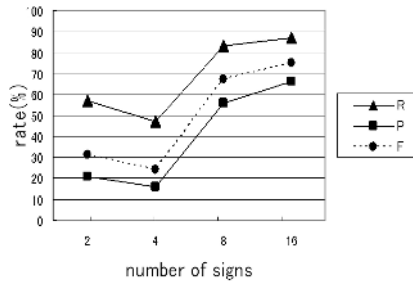


Fig. 7. Experiment B: R , P , and F within 60 sec.

5 Conclusion

We proposed a method of retrieving semantically similar scenes to a query video segment with sign sequence matching based on audio features. We tried to cluster the extracted audio feature vectors by K-means clustering algorithm and assign the signs, which are the cluster IDs, to the feature vectors in the corresponding clusters. The similar scenes were retrieved by sign sequence matching based on dynamic programming. As a result of applying the proposed method to entertainment programs and news programs, when the errors are within 60 seconds and the number of signs are eight or sixteen, we were able to obtain the similar scenes with the F-measure of 70% within 1 second.

Successful retrieval can be realized when each scene has different audio characteristics. However, when several types of scenes share the same characteristics, there will be more false detection. In order to solve this problem, developing an efficient way to combine visual features which maintains the short processing time will be our future work.

References

1. H. Sundaram and S.-F. Chang: "Video Scene Segmentation Using Video and Audio Features," Proc. IEEE ICME 2000, July 2000.
2. Y. Wang, Z. Liu, and J. Huang: "Multimedia Content Analysis Using Both Audio and Visual Clues," IEEE Signal Processing Magazine, pp.12-36, Nov. 2000.
3. Y. Deng and B. S. Manjunath: "Content Based Search of Video using Color, Texture and Motion," Proc. IEEE ICIP, Vol 2, pp.13-16, Oct. 1997.
4. M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan: "Multi-Modal Extraction of Highlights from TV Formula 1 Programs," IEEE ICME, Aug. 2002
5. Tong Zhang and C.-C. Jay Kuo: "Audio Content Analysis for On-line Audiovisual Data Segmentation," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 4, pp.441-457, May 2001.
6. Z. Liu, J. Huang, and Y. Wang: "Classification of TV Programs Based on Audio Information Using Hidden Markov Model," Proc. IEEE 2nd Workshop on Multimedia Signal Processing, pp.27-32, Dec. 1998.
7. N. Patel and I. Sethi: "Audio Characterization for Video Indexing," Proc. SPIE, vol.2670, pp.373-384, 1996.
8. K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura: "Video Handling with Music and Speech Detection," IEEE Multimedia, pp.17-25, Fall 1998.
9. E. Scheirer and M. Slane: "Construction and Evaluation of a Robust Multifeature Speech / Music Discriminator," Proc. ICASSP 97, vol.II, pp.1331-1334, 1997.

Architecture and Analysis of Color Structure Descriptor for Real-Time Video Indexing and Retrieval

Jing-Ying Chang, Chung-Jr Lian, Hung-Chi Fang, and Liang-Gee Chen

DSP/IC Design Lab.,
Graduate Institute of Electronics Engineering,
National Taiwan University, Taipei, Taiwan
{jychang, cjlian, honchi, lgchen}@video.ee.ntu.edu.tw
<http://video.ee.ntu.edu.tw>

Abstract. Color structure descriptor (CSD) provides satisfactory image indexing and retrieval results among other color-based descriptors in MPEG-7. The superiority comes from the consideration of space distribution of pixel colors. In this paper, we proposed the first CSD hardware architecture which can generate CSD description with frame size 256×256 and 30 frames per second (fps). This architecture provides about 12 times speed-up than running on a 2.54 GHz microprocessor platform to achieve real-time applications like assisting rate control in video coding system and circumstance change detection in surveillance system.

1 Introduction

With mature digital video technology, inexpensive camcorders gradually enter our life. More and more multimedia are produced and shared among the world. Original intention of MPEG-7 is to provide a powerful search engine which helps people easily find what they are looking for. Some MPEG-7 toolkits further integrate useful functionalities for categorizing and organizing their personal collection. However, some related research [1] showed that most people only categorize their albums at semantic level, but the recognition technique nowadays is still not able to meet this kind of demand. MPEG-7 descriptors are good tools for indexing and retrieval but should not be limited to them. MPEG-7 descriptors can be creatively extended and linked to applications such as rate control in real-time video coding and movement detection in surveillance systems. In these applications, computational loads of the real-time implementation for these descriptors will not be a trivial issue.

With statistics derived from MPEG-7 descriptors, good indication of image and video properties can provide referable adjustment parameters for video pre-processing like auto white balance, RGB gains tuning, saturation control, auto contrast, and edge enhancement. In video coding, it can assist fast algorithm of motion estimation, rate control policy, probability distribution model of entropy

coding, and so on. When we use them in surveillance system, the system can notice police to keep an eye on unusual behavior by analyzing object trajectory. Face descriptor can also provide auto identification of uncertified people in certain degree.

MPEG-7 visual descriptors record statistics of images and video sequences in color, texture, shape of objects, and motion. Because the variety of possible applications, we first take implementation of color descriptors as our start point. Color is one of important visual attributes for human vision and image processing. It is also an expressive visual feature in image and video retrieval. Color descriptions usually are irrelevant to viewing angle, translation and rotation. This advantage possesses good resistance to undesired shaking of camera. In MPEG-7, six descriptors are selected to record color statistics of images and video. Among them, CSD provides best image indexing and retrieval results[2]. The superiority comes from that CSD considers space distribution of pixel colors by recording appearance of each color in every structuring element window in its histogram [3]. In this paper, we focus on the architecture and analysis of CSD.

The challenge to realize CSD hardware accelerator for real-time video system is that each pixel in a frame needs to be scanned 64 times. The vast data bandwidth and then excessive operating frequency make CSD impossible for real multimedia systems. Analysis of the trade-off between input bandwidth and local buffer size is the first issue needed to be evaluated. Then, the index algorithm of the color appearance in one structuring window (SW) has to be considered carefully to lower operating frequency. Along with exploring suitable solutions, hardware extensibility should not be left behind. It is worth to integrate with other descriptors with small overhead.

Operational analysis of software simulation is shown in Table 1. "Accumulation" comprises related operations of moving SW and CSD histogram accumulation. For a video sequence with frame size 256×256 , 30 fps, 4.5 giga instructions per second (GIPS) and 6 giga bytes per second (GB/s) of memory bandwidth are required in one second. Such computational cost is the reason why CSD can not be applied to real-time products without a hardware accelerator. And there is no good solution at present.

In this paper, we first describe briefly the algorithm of CSD in section 2. Before going into implementation details of each functional block, hardware issue and operational parallelism are discussed in section 3.1 and then each block design. Section 4 shows the experimental result and summarizes the chip specification. Section 5 is dedicated to concluding remarks and future research.

2 Color Structure Descriptor

CSD represents an image by color accumulation and the local spatial distribution of colors. The procedure of CSD histogram uses a 8×8 SW to observe which colors are presented in it, and then updates those color bins by only adding one, no matter how many same color pixels exist. Figure 1 shows that two images have different CSD description with the same traditional histogram[4]. Right

Table 1. MIPS and memory bandwidth of CSD generator.

Operation	1 fps		30 fps	
	Number of instructions (MIPS)	Memory bandwidth (MBytes)	Number of instructions (MIPS)	Memory bandwidth (MBytes)
HMMD	5.625	3.585	168.750	107.550
Accumulation	143.657	202.456	4309.710	6073.680
Quantization	0.051	0.001	1.517	0.039
Others	0.990	0.697	29.713	20.901
Total	150.323	206.739	4509.690	6202.170

image looks more scattered than left one. Such situation causes gray pixels exist in more SWs and reflects on gray bin in CSD description. This advantage let us easily distinguish those images with similar dispersion.

Figure 2 depicts CSD extraction procedure [5]. Our design chose highest number of bins for more precise CSD description in real-time applications. The top path directs the flow of 256-bin CSD. It starts with color transformation from RGB to HMMD. Next step is histogram accumulation which is followed by a decision of number of bins needed. After a nonlinear quantization, CSD description is derived.

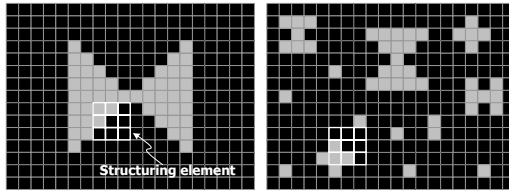


Fig. 1. Two images have the same traditional histogram, but right one has much more gray components in CSD description.

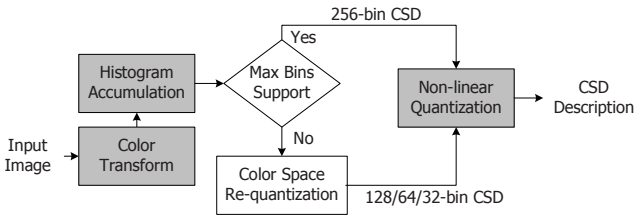


Fig. 2. CSD extraction flow.

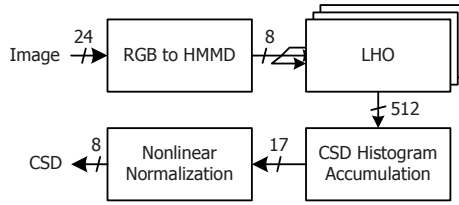


Fig. 3. Block diagram of CSD architecture.

3 Computational Complexity and Proposed Architecture

As described in Section 1, we focus on real-time applications of MPEG-7 like video coding assistance and surveillance systems. Besides, generated CSD descriptions still can be used for search of multimedia contents. And for supporting comparison with descriptions generated by other tools, 256 levels of color quantization is adopted for downscale comparison.

Since a sub-sample factor is defined in the standard for large images, we choose 256×256 as input image size. The sub-sample factor, K , is defined as $K = \max\{1, 2^{\lfloor \log_2 \sqrt{W \cdot H} - 7.5 \rfloor}\}$, where W and H are the width and height of image. For example, $K = 2$ implies an image is sub-sampled by 2 horizontally and vertically. Note that the SW size is always 8×8 .

Our CSD block diagram is shown in Fig. 3. After color transformation, pixels are sent to corresponding local histogram observing (LHO) blocks and index colors that exist in these windows. Summation of outputs of three LHO blocks indicates how many windows does each color belong in. Then, the summation updates CSD histogram. Finally, CSD description is obtained via non-linear quantizing the completed histogram.

3.1 Parallelism Analysis

Specification of our CSD generator is for the video sequence with frame size 256×256 and 30 fps. Operating frequency limitation is targeted at 27 MHz, which is common for most TV systems. This requirement can be achieved by buffering three successive SWs (8×10 pixels). Purpose of the buffer is for data sharing. The scan order is shown in Fig. 4. Pixel values of three SWs are complete updated after discarding top row pixels from last three SWs and reading in ten new bottom pixels in current SWs. After finishing indexing SW colors in one stripe, we start to index SW colors in next stripe. The displacement between adjacent stripes is three pixels.

Parallelism decision is according to the target frequency. Approximately, in the situation of no local buffer of SW, each pixel in every window has to be scanned again even though it has been scanned during the period of operations of last neighboring window. The memory bandwidth is about 357 megabytes/sec (MB/s) and the required operating frequency is 119 MHz. In fact, we assume

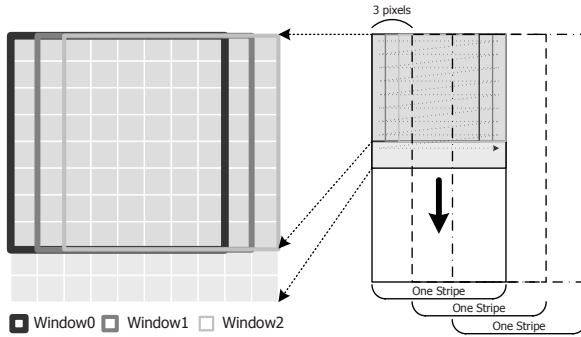


Fig. 4. Pixel scan order of three structuring windows.

Table 2. Relationship between parallelism and operating frequency. Zero parallelism means no SW is buffered. The minimum requirement to meet target frequency (27 MHz) is three parallelism.

Parallelism	MB/s	MHz
0	357	476
1	46	61
2	26	35
3	19	25
4	16	21

histogram can be updated once in one cycle to make this chip running at 119 MHz. But according to the problem described in section 3.2, it takes four cycles to update one pixel data on average and forces the required operating frequency to 476 MHz. Relationship of parallelism and operating frequency is shown in Table 2. Three parallelism is the final decision to meet the requirement without over design.

3.2 Color Appearance Recording in LHO

How to record which colors exist in a SW efficiently is another main issue. It is unrealistic to query all pixels at the same time or to query by taking 64 cycles. The method of querying at the same time will make interconnection of decision circuit become very large and inconvenient to handle. The method of querying by taking 64 cycles has to be realized by raising operating frequency. In order to solve the problem, we proposed a LHO architecture. LHO contains a SRAM to record color histogram of a SW and a color appearance register bank to indicate which colors exist in the SW according to the values of the color bins.

The main idea of LHO is recording SW histogram to indicate which colors exist in a SW. Along with updating histogram, we observe the value of changing color bin and save this information into color appearance register bank. Nonzero

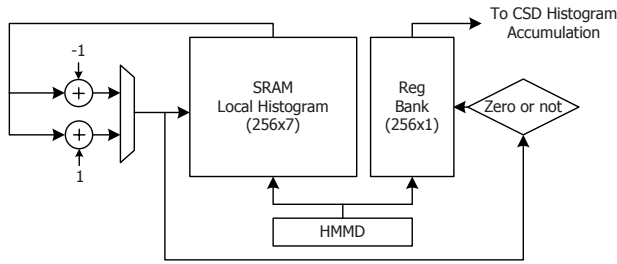


Fig. 5. Structuring window histogram updating architecture.

bin means this color belongs to the window. After update, three register banks are summed and sent to CSD histogram accumulation block.

Using SRAM to record histogram of SW is an area efficient method. But histogram updating cycles are directly restricted by SRAM specification. Single port SRAM provides one read or one write in a cycle. That means, when we get an address from the color which needs to update corresponding color bin, we read the bin value in one cycle, add or subtract the value by 1, and write it back to SRAM in another cycle. With an appropriate design for dual port SRAM, the throughput of updating histogram can achieve one update per cycle at the expense of double SRAM area and power. With power consideration, we choose single port SRAM as buffer of SW histogram. Single port SRAM takes four cycles to refresh histogram for each pixel. Two cycles are for removing accumulation from previous pixel and the others are for addition of incoming pixel. To update three SWs by refreshing ten pixels will take 40 cycles. Figure 5 shows the LHO architecture.

3.3 CSD Histogram Accumulation

According to cycle analysis in section 3.2, there are 40 cycles to complete 256-bin CSD histogram accumulation. If we store CSD histogram in SRAM and wish to refresh it in time, that means we have to update 16 color bins in two cycles, improper bit-width and number of addresses of SRAM will cause this SRAM occupies large area and waste much power. Here we divide this SRAM into four to lower the unreasonable bit-width. The bit-width is equal to four color bins. Each SRAM is 16×64 bits.

3.4 Non-linear Quantization

After CSD histogram accumulation is finished, non-linear histogram quantization is the final step. Each bin should be quantized into 8-bit via 255 comparisons. With binary comparison method and folding skill, eight comparisons are needed to quantize one bin. This strategy is shown in Fig. 6. As shown in (a), we compare the bin with center value of valid range each time. Since the latency of non-linear quantization, which is compared with CSD histogram accumulation,

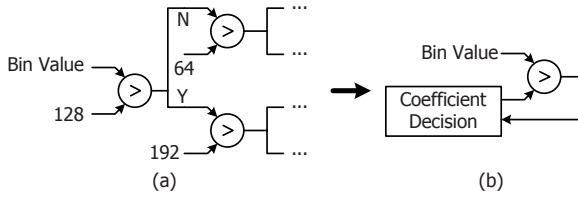


Fig. 6. Folding skill on non-linear quantization.

Table 3. Indexing and retrieval result

Descriptor	Color space	ANMRR
CSD	HMMD	0.00105097
CSD	YCbCr	0.00360790
SCD	HSV	0.00165656
SCD	YCbCr	0.00428604

Table 4. Chip specification

	Technology	UMC 0.18 μm CMOS 1P6M
	Core size	$1.36950 \times 1.36584 \text{ mm}^2$
	Gate count	49865
On-chip single port SRAM		11136 Bits
	Max frequency	31.25 MHz
	Operating frequency	27 MHz
	Processing speed	256×256 , 30 fps@ 27 MHz
	Power Consumption	39.53 mW@ 27 MHz, 1.8 V

is negligible, 255 comparators can be folded into one. With (b) architecture, 2048 (256×8) cycles and one comparator are needed to achieve this work.

4 Experimental Result

Our indexing and retrieval database contains 526 images in 78 categories. Those images are collected from Internet and manual categorized. Furthermore, for extending the concepts of these descriptors to image and video coding, we replace default color spaces with YCbCr domain and the performance drops slightly.

Here we use a quantitative measure method called query-by-example (QBE) suggested by MPEG-7 [4]. QBE sorts the distances between description vector of query image and those of images contained in a database. The smaller average normalized modified retrieval rank (ANMRR) means the descriptor provides better indexing and retrieval ability.

Table 3 shows the indexing and retrieval results of CSD and scalable color descriptor (SCD) with designated and YCbCr color spaces. SCD listed here is for comparison. The results with YCbCr are also acceptable and imply that we can apply the concepts to the field of image and video coding which chooses YCbCr as default color space.

This first proposed CSD hardware architecture for realtime applications can generate CSD description with frame size 256×256 @ 30 fps. Detailed specification is shown in Table 4.

5 Conclusion

In this paper, we provide the vision of future MPEG-7 descriptor applications for not only indexing and retrieval, but also for real-time multimedia applications. First analysis of dedicated hardware architecture design for MPEG-7 CSD descriptor is also proposed. Detailed design explorations of the hardware implementation, and practical reference data of prototype is valuable for future researchers. The integration with SCD by sharing much existed resource is ongoing. In the future, descriptors with similar architecture can be integrated into this design.

References

1. Kerry Rodden, Kenneth R. Wood: How Do People Manage Their Digital Photographs. Proceedings of the conference on Human factors in computing systems, ACM Press, New York, NY, USA (2003) 409–416
2. Ojala, T. and Aittola, M. and Matinmikko, E.: Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories. IEEE International Conference on Pattern Recognition, 2002, Vol. 2 (August 2002) 1021–1024
3. Qian, R.J. and Van Beek, P.J.L. and Sezan, M.I.: Image retrieval using blob histograms. IEEE International Conference on Multimedia and Expo, 2000, Vol. 1 (August 2000) 125–128
4. B.S. Manjunath and Philippe Salembier and Thomas Sikora: Introduction to MPEG-7. JOHN WILEY and SONS (2002) 204–208
5. Leszek Cieplinski and Munchurl Kim and Jens-Rainer Ohm and Mark Pickering and Akio Yamada: Text of ISO/IEC 15938-3/FCD Information technology - Multimedia content description interface - Part 3 Visual, ISO/IEC JTC 1/SC 29/WG11 N4062 (March 2001) 47–52

Automatic Salient-Object Extraction Using the Contrast Map and Salient Points

SooYeong Kwak¹, ByoungChul Ko², and Hyeran Byun¹

¹ Dept. of Computer Science, Yonsei University, Seoul, Korea, 120-749
{ksy2177, hrbyun}@cs.yonsei.ac.kr

² Telecommunication R&D Center, SAMSUNG ELECTRONICS CO., LTD, Korea
byoungchul.ko@samsung.com

Abstract. In this paper, we propose a salient object extraction method using the contrast map and salient points for object-based image retrieval. In order to make the contrast map, we generate three-feature maps such as luminance map, color map and orientation map and extract salient points from an image. By using these features, we can decide the Attention Window (AW) location easily. The purpose of the AW is to remove the useless regions included in the image such as background as well as reducing the amount of image processing. To create the exact location and flexible size of the AW, we use above features with some proposed rules instead of using pre-assumptions or heuristic parameters. After determining of the AW, we apply the image segmentation to inner area of the AW and combine the candidate salient regions as one salient object.

1 Introduction

Region-Based Image Retrieval (RBIR) research is a promising research field because it exploits regions such as a face, a body of car, the sun instead of using low-level features such as color, texture as keys to retrieve images. However, since humans are accustomed to utilizing high-level concepts, for example, human, car, flower rather than the separated regions, RBIR has many limitations for retrieving semantic objects. Therefore, it is an essential work to extract salient object from an image for semantic image retrieval.

Salient object extraction method can be used not only the image retrieval, but also many other applications. First, it can be used for suitable representation of images in various display devices such as a cellular phone and a PDA (Personal Digital Assistant) [1]. That is, if someone wants to send an image that includes a bird in a bush, from PC to the smaller display device such as a cellular phone, that image has to be reduced to the size of a screen of a cellular phone regardless of the contents of an image. Therefore, the receiver may not be able to figure out the meaning of an image. In this case, if the salient object is extracted and it can be cropped as the proper size of the screen of a cellular phone before sending, it helps the user figures out the meaning of an image.

Second, it can be used in image watermarking system [6]. Most existing watermarking techniques embed watermarks in the entire image without considering

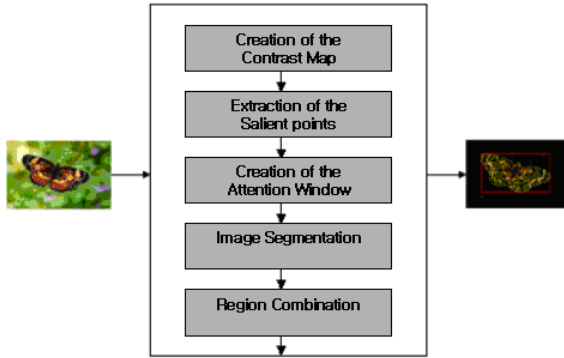


Fig. 1. Flow chart on salient object extraction.

the image object. However, generally people tend not to transform the salient object. When watermarks are embedded into the entire image, the watermarks would be severely destroyed by cropping of the background regions. Therefore, it is more effective that watermarks are embedded into the salient object instead of the entire of an image.

Many researches are already proposed for extraction of the salient object from an image. Kim et al. [2] assumed that salient object is generally located near the center of the image after that they automatically extract the salient object in an image. However, salient object is not always located near the center of image. Wang [7] and Osberger [8] use some features and heuristic parameters for features in order to extract the salient object. However this is not automatic method because the user has to control the optimal parameters.

In this paper, we propose a new method for extraction of salient region using the contrast map and salient points automatically. Fig 1 shows the flow chart of our proposed method. The rest of this paper is organized as follows: Section 2 and section 3 introduces the concept of the contrast map and the salient points. In Section 4, detection of the AW is described. Section 5 shows the evaluation results obtained from user study experiment. The conclusion and future works are presented in section 6.

2 Contrast Map

We decide the initial AW location according to saliency of the contrast map. Before the extraction of the AW, we make the contrast map which has getting location information. The contrast map uses three features that are luminance, color and orientation. These three features make the each feature map. In this paper, we modified the Itti's visual attention model [3] to create the contrast map. Moreover, we prove that color map may hurt the object extraction performance compare to other feature maps in case the color of salient object is similar with the background. The experimental results are shown in section 5.

For contrast map, we create six luminance maps, twelve color maps and eighteen orientation maps.

2.1 Luminance Map

Before creating the luminance map, we apply the gaussian filter to the image in order to remove noises. Then, we apply the different size of filters to the down-sampled image. The filter estimates difference between a "center-point" and surrounded-points of image scale c and filter scale s . In Equation (1), $G(c,s)$ is the intensity difference between a center point and surrounded points. In this method, the feature map is sensitive to the size of contrast-filters. If the size of an object is bigger than the filter size, saliency of luminance map has low contrast value and if it has smaller than the filter size, saliency of luminance map has high contrast value. To overcome this defect, we reduce the image size by a factor of 2, 3 and 4. The filters are also adjusted by the factor of the image size like Equation (1). By using the Equation (1), we can make six feature maps. To combine the different size of feature maps with one image, we normalize them to one quarter of the original image size. In Equation (1), $D(\cdot)$ is the combined feature map at scale of c . Six maps are all summed and normalized into luminance map. Fig. 3-(b) shows three examples of luminance map.

$$\bar{L} = \frac{1}{6} \sum_{c=2}^4 D\left(\sum_{s=c+3}^{c+4} G(c,s)\right) \quad (1)$$

2.2 Color Map

In order to get color map, we use the CIE $L^*a^*b^*$ color model because it is a device independent and it is similar to human color perception. Here, we use the a^* and b^* color model except the L^* (luminance of $L^*a^*b^*$). The process of color map is similar to the process of luminance map. As the same method with the luminance map, we use the different size of filters and images. We create six feature maps from each a^* and b^* color model and normalize them as one color feature map. Before creating color feature map, we also apply the gaussian filter to the image. In Equation (2), $A(c,s)$ and $B(c,s)$ is the color difference between a center point and surround points at image size c and filter size s . Fig. 3-(c) shows three examples of color map.

In general, color map is known as a major feature to distinguish salient and non-salient object. However, in this paper, we prove that the color map is not always good feature for extraction of the salient object. Especially, incase the salient object has protective coloration, color map rather hurt the extraction performance. Fig. 2 shows some examples of exceptional cases of color map and the experimental results are shown in section 5.

$$\bar{C} = \frac{1}{12} \left(\sum_{c=2}^4 D\left(\sum_{s=c+3}^{c+4} A(c,s)\right) + \sum_{c=2}^4 D\left(\sum_{s=c+3}^{c+4} B(c,s)\right) \right) \quad (2)$$



Fig. 2. Four exceptional cases of color map. In this case, color map does not play an important role to extract salient object and rather it may hurt the extraction performance.

2.3 Orientation Map

Color is based on the observation often is used to encode functionality (sky is blue, forests are green) and in general will not allow us to determine an object's identity [8]. Therefore, texture or geometric properties are needed to identify objects. Since salient regions have special texture information, we use the wavelet transform in order to get texture information at different scales. After the one level wavelet transform, we get the horizontal (HL), vertical (LH) and diagonal (HH) orientation information from the wavelet subbands. Then, we create the orientation map as the same method with luminance and color map. In Equation (3), $O(c, s)$ is the coefficient difference between a center coefficient and surrounded coefficients at image size c and filter scale s . From Equation (3), We can get the eighteen feature maps and these are normalized into one orientation map. Fig 3-(d) shows three examples of orientation map.

After creation of three feature maps, they are combined as one contrast map. Fig. 3-(e) shows three examples of the final contrast map.

$$\bar{O} = \frac{1}{18} \left(\sum_{HH,HL,LH} \left(\sum_{c=2}^4 D \left(\sum_{s=c+3}^{c+4} O(c, s) \right) \right) \right) \quad (3)$$

3 Salient Point

In order to create accurate AW, we extract the salient points as well as contrast map. Salient points are a set of "interesting points" within an image that exist in both a corner-like area and a high signal variation occurrence area. We use a wavelet based salient points detection method [5] to extract salient points where variations occur in an image. In this method, we do not fix the number of salient points in order to reflect the variation of image content. Instead, we set the fixed threshold to exclude the low saliency value according to image content. In Fig. 4, the yellow points are salient points.

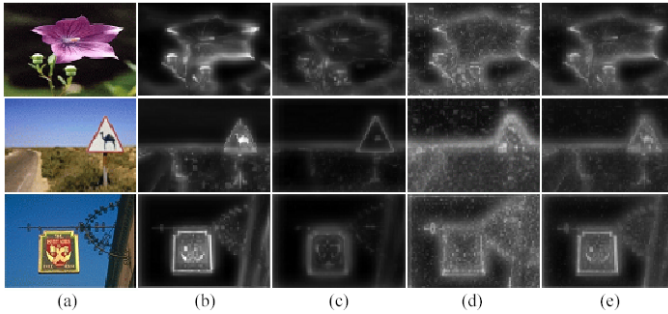


Fig. 3. (a) Original image (b) Luminance map (c) Color map (d) Orientation map (e) Contrast map

4 Create the Attention Window and Segmentation

After the contrast map is created and salient points are extracted, the AW should be detected. The purpose of used AW is to remove the useless regions included in the image such as background and to reduce the amount of image processing for region segmentation. Therefore, the AW size itself is important. In this paper, we determine the proper size of the initial AW as the three quarters of the image and reduce the size until it meets the predefined conditions. First, in order to determine the proper location of the AW, we choose one window that has the maximum number of salient points in the full image and the highest saliency part in contrast map. Then, we reduce the size of the AW as the approximate size of salient object. To do this, we repetitively reduce the size from original AW until the AW meets the optimal condition. That is, the value of contrast map and the number of salient point should over predefined thresholds. Fig. 4 shows the final AW.

If we detect the optimal position and size of the AW, we segment the inner area of the AW. We use the FRIP segmentation method in this system [4]. After that, we first remove the outside regions of the AW. Then, additional two constraints are applied to the candidate salient regions in the AW. First, if the candidate regions should exist in the AW. If the number of pixels of a region, which exists outside of the AW, is over one third of the overall number of its pixels in the region, we remove it from the candidate salient regions. Second, if the boundary length of a region, which connects with the border of the image is over one third of the overall length of its boundary, we also remove it. After the final salient regions are determined, these regions are combined as one salient object. Fig. 5 shows the result of the extraction of the salient object.

5 Experimental Results

To show the soundness of our algorithm, we test the extraction performance of the salient regions. To complete this experiment, we randomly chose 100 im-



Fig. 4. The final Attention Window

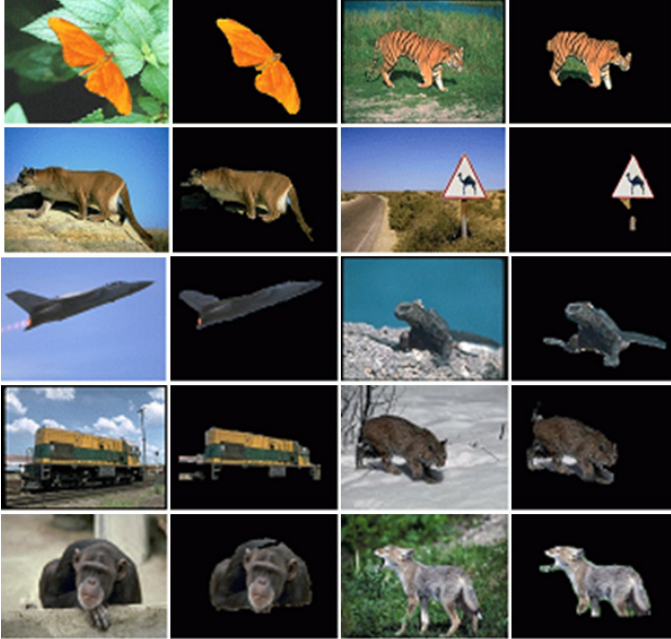


Fig. 5. Result of salient object extraction.

ages from the Corel-photo CD, covering a wide variety of content ranging from natural images to graphic images without pre-selected categories. The randomly selected images have specific objects in the images. Then we cropped the salient object using the graphic tool from each image and compared them to the extraction results of our system. Because there is no specific method to evaluate the performance of our algorithm, we modify the evaluation method (Equation (4)) that is proposed by Kim et al. [5].

$$S_U = (M - (M \cap E))/S_M, S_O = (E - (M \cap E))/S_E \quad (4)$$

$$A = 100 - (S_U + S_O) \times 100$$

The S_M and S_E represent the cardinality of M and E , respectively, where M represents the set of pixels in the manually extracted central object and E represents the set of pixels in the extracted salient regions. The symbols S_U and S_O and represent inaccuracy of under-extraction and of over-extraction,

Table 1. Performance comparison between two methods of the contrast map

	S_U	S_O	A
L+O	29.7	16.3	56.4
L+C+O	27.2	21.9	53.9

Table 2. Performance comparison between three methods

	S_U	S_O	A
Salient point	29.8	14.1	59.2
Contrast map	29.7	16.3	56.4
Salient point+Contrast map	30.8	11.4	61.1

respectively. The symbol A represents the accuracy between central object and salient regions.

In this paper, we perform two experiments in order to compare the performance of our method. First, we only use contrast map to create the AW without using the salient points. To back up our point of view which color is not always signalize the salient object, we make two feature maps. The first one is made by the combination of luminance and orientation map (L+O) and the second one is made by the combination of three feature maps including color map (L+C+O).

The table 1 shows the experimental results evaluated by Equation 4. It shows that combination of two feature maps is better performance than combination of three feature maps. This result indicates that the color is not always pivotal factor for human perception because some salient objects have protective coloration against the background.

The second, to evaluate the function of the AW, we compare the three methods for determining the AW. First, we only use the salient points to make the AW and second we also only use the contrast map. Third, we use both contrast map and salient points to make the AW. From three AWs, we extract salient object and evaluate the performance by using Equation (4). The table 2 shows the performance among three methods. It shows that the above two methods supplement with each demerit. That is, when the AW cannot detect the accurate position of a salient object just using the contrast map, the salient point helps to detect accurate position of the salient object. Also when the salient point scattered around all images, the contrast map helps to make the AW. In other words, the contrast map and salient point contribute to make the accuracy of the AW position and the size of the AW.

6 Conclusion

We presented a novel method for extracting salient object. Unlike the previous methods, we used both contrast map and salient points to extract accurate salient object and proved that color map hurt the extraction performance com-

pared to other feature maps. Experimental results showed that the proposed method could extract meaningful salient object well without any assumption and passive techniques. The proposed framework can be easily employed or integrated into a variety of vision systems and visual content analysis related multimedia applications.

Acknowledgements. This work was supported (in part) by the Ministry of Information & Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.

References

1. Ma Y.F. and Zhang H.J.: Contrast-based Image Attention Analysis by Using Fuzzy Growing. ACM Int. Conf. on Multimedia. (2003) 355-358
2. Kim S., Park S. and Kim M.: Central Object Extraction for Object-Based Image Retrieval. Int. Conf. on Image and Video Retrieval (2003) 39-49
3. Itti L., Koch C. and Niebur E.: A Model of Saliency-based Visual Attention for Rapid Scene Analysis. IEEE Trans. on PAMI. **20**. (1998) 1254-1259
4. Ko B.C. and Byun H.R.: Region-based image retrieval: A new method for extraction of salient regions and learning of importance scores. Int. Journal of Patt. Recog. and Arti. In-telli. **17** (2003)
5. Loupias E. and Sebe N.: Wavelet-based Salient Points for Image Retrieval. Research Report RR 99.11. RFV-INSA Lyon. (1999)
6. Lei G. and Bao-long G.: A Watermarking Scheme Using Image Object Region. IEEE Int. Conf. on Computational Intelligence and Multimedia Applications **5** (2003) 419-423
7. Wang W., Song Y. and Zhang A.: Semantics retrieval by region saliency. Int. Conf. on Image and Video Retrieval (2002) 29-37
8. Stricker M. and Dimai A.: Spectral covariance and fuzzy regions for image indexing. SPIE-Storage and Retrieval for Image and Video Database (1997)

Shape-Based Image Retrieval Using Invariant Features

Jong-Seung Park¹ and TaeYong Kim²

¹ Department of Computer Science & Engineering, University of Incheon, 177 Dohwa-dong, Nam-gu, Incheon, 402-749, Republic of Korea
jong@incheon.ac.kr

² Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, HukSuk-dong 17, DongJak-gu, Seoul, 156-756, Republic of Korea
kimty@cau.ac.kr

Abstract. In this paper we propose an accurate shape-based image retrieval method which uses both types of moment invariants and Fourier descriptors. The method first excludes irrelevant images of different appearances using the region-based moment invariant features and then, using Fourier descriptors, the method increases the retrieval effectiveness substantially. The retrieval method was tested on our shape databases and the search accuracy was compared with those of Fourier descriptors and moment invariants. The hybrid method of using both boundary-based Fourier descriptor and region-based moment invariants provides much better performance than other similarity method.

Keywords: Shape retrieval, Fourier descriptors, Moment invariants.

1 Introduction

Due to inexpensive digital imaging technologies and wide use of digital cameras, very large digital image archives have been created and used in numerous applications. Together with the increase in the number of image archives, there has been rapid progress in visual information indexing and retrieval methodologies [1][2]. The useful visual attributes include color, texture, shape and location. Among them, shape is one of key visual features for distinguishing visual data. This paper is concerned with shape descriptors which measures perceptual similarity of different shapes. In many shape-based image search and retrieval systems, moment invariants and Fourier descriptors are considered as the most representative measures in 2D shape matching [3][4]. Theoretically, both types of measures keep invariant properties to scale change and rotation in 2D space. In a digital image, the invariant properties do not generally hold due to the generic problem of the discrete function.

To overcome the numeric problem we introduce a two-stage similarity scheme. The basic idea of the proposed method is that the moment invariants are specially effective to the classification of appearances and the Fourier descriptors

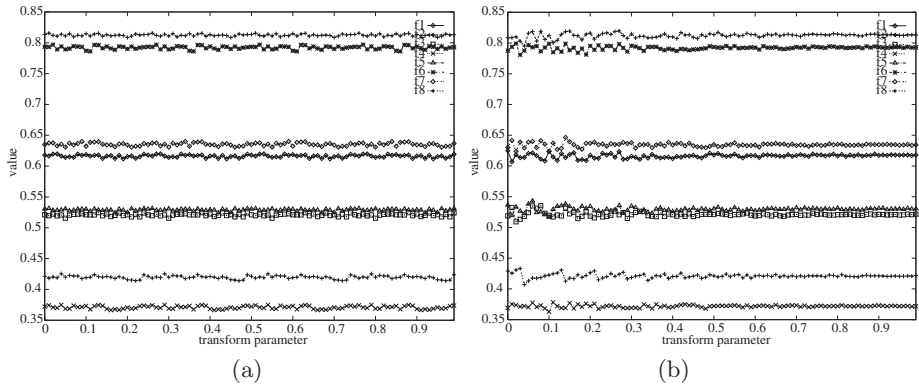


Fig. 1. Invariant properties of the Fourier descriptor feature vector \mathbf{f}_F : (a) \mathbf{f}_F under rotation, (b) \mathbf{f}_F under scaling.

to the discrimination of object poses. The method first excludes irrelevant images of different appearances using the region-based moment invariant features. Then the method orders the filtered images according to the similarity of the given pose. To rank images according to the pose similarity we compute Fourier descriptors. This step increase the retrieval effectiveness substantially.

Section 2 and Section 3 describes invariant properties of Fourier descriptors and moment invariants, respectively, that are used in our method as similarity measures for image retrieval. Then, in Section 4, the proposed similarity scheme which uses both moment invariants and Fourier descriptors is described. Experimental results are presented in Section 5, and conclusion and future works are offered in Section 6.

2 Fourier Descriptors

Fourier descriptors are 2D invariant features available from boundary points. Suppose that the boundary of a particular shape has N pixels numbered from 0 to $N - 1$ and the contour is described as two parametric equations:

$$x(k) = x_k \text{ and } y(k) = y_k, \text{ where } k = 0, \dots, N - 1.$$

By considering the equations in the complex plane, a direct parametric representation $z(k)$ is possible:

$$z(k) = x(k) + jy(k) .$$

The Fourier descriptors $Z(t)$ of the curve is the discrete Fourier transform coefficients of the complex valued curve $z(k)$:

$$Z(t) = \frac{1}{N} \sum_{k=0}^{N-1} z(k) \exp\left(\frac{-j2\pi kt}{N}\right) .$$

A simple normalization of $Z(t)$ makes the Fourier descriptors invariant to the starting point of sampling, rotation, scaling and translation. Each coefficient of a Fourier descriptor has two components, amplitude and phase. By using only the amplitude component, we achieve rotation invariance as well as the invariance to the starting point. By dividing all amplitudes by the amplitude of the first non-zero frequency coefficient, we achieve the scale invariance. Since only the DC coefficient is dependent on the position of shape, it is discarded to achieve the translation invariance. We compute the m -dimensional feature vector \mathbf{f}_F from m Fourier descriptors by dividing the magnitudes by $|Z(1)|$:

$$\mathbf{f}_F = \left(\frac{|Z(-m/2)|}{|Z(1)|}, \dots, \frac{|Z(-1)|}{|Z(1)|}, \frac{|Z(2)|}{|Z(1)|}, \dots, \frac{|Z(m/2 + 1)|}{|Z(1)|} \right). \quad (1)$$

In our system, we choose $m = 16 (= 2^4)$ so that the transformation can be conducted efficiently using FFT.

3 Moment Invariants

Moment invariants are useful in 2D object recognition. Moment invariants are functions of moments that are invariant under certain transformations. Although, moments are defined on a continuous image intensity function, a simple approximation is possible for a discrete binary image using a summation operation. Let f be a binary digital image matrix with dimension $M \times N$, and let $S = \{(x, y) | f(x, y) = 1\}$ represent a 2D shape. The moment of order (p, q) of shape S is given by $m_{pq}(S) = \sum_{(x,y) \in S} x^p y^q$. The central moment of order (p, q) of shape S is given by $\mu_{pq}(S) = \sum_{(x,y) \in S} (x - \bar{x})^p (y - \bar{y})^q$ where (\bar{x}, \bar{y}) is the center of gravity: $\bar{x} = m_{10}(S)/m_{00}(S)$, $\bar{y} = m_{01}(S)/m_{00}(S)$. From the central moments, the normalized central moments, denoted by η_{pq} , are defined as $\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}$ where $\gamma = \frac{p+q}{2} + 1$ for $p + q = 2, 3, \dots$.

From the second- and third-order normalized central moments, a set of seven invariant moments, which is invariant to translation, scale change and rotation, has been derived by Hu [5][6][7]:

$$\begin{aligned} f_{H1} &= \eta_{20} + \eta_{02} \\ f_{H2} &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ f_{H3} &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ f_{H4} &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ f_{H5} &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right] \\ &\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right] \\ f_{H6} &= (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ f_{H7} &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right] \\ &\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right] \end{aligned}$$

Hu [6] has proved the invariance properties of the seven moments for the case of continuous functions. The seven moments constitutes the Hu feature vector \mathbf{f}_H :

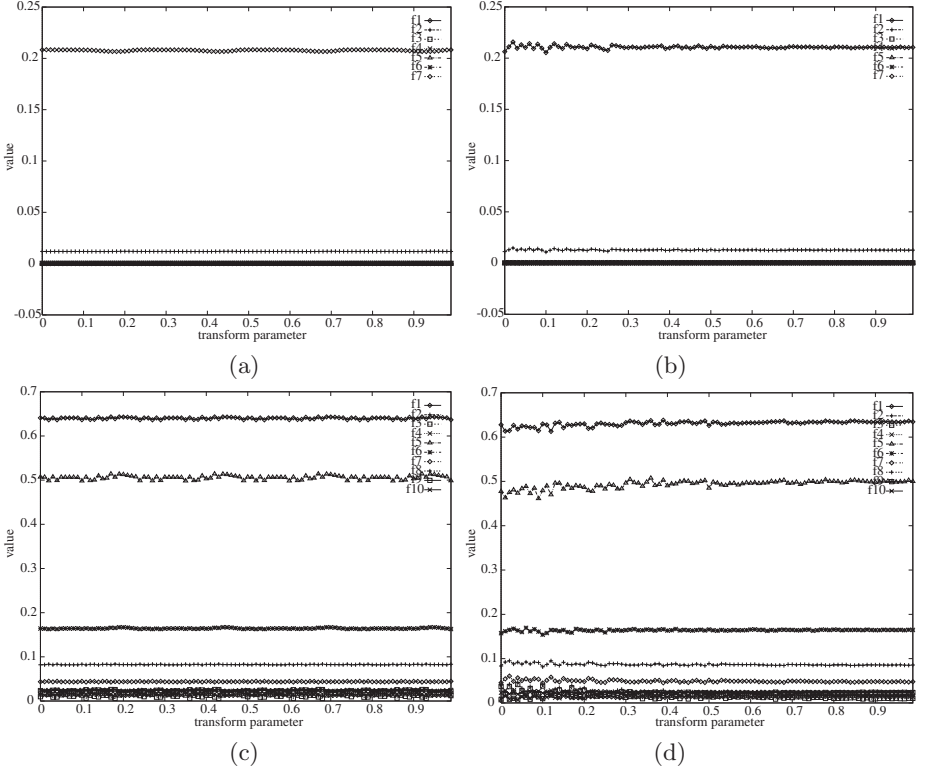


Fig. 2. Invariant properties of the vector \mathbf{f}_H for the Hu moment invariants and the vector \mathbf{f}_Z for the Zernike moment invariants: (a) \mathbf{f}_H under rotation, (b) \mathbf{f}_H under scaling, (c) \mathbf{f}_Z under rotation, (d) \mathbf{f}_Z under scaling.

$$\mathbf{f}_H = (f_{H1}, \dots, f_{H7}). \quad (2)$$

The Zernike moment of order n with repetition m that vanishes outside the unit circle is

$$A_{nm} = \frac{n+1}{\pi} \sum_{(x,y) \in S, x^2+y^2 \leq 1} V_{nm}^*(\rho, \theta)$$

where $R_{nm}(\rho)$ and V_{nm}^* are defined in [8]. The magnitudes of the Zernike moments, $|A_{nm}|$, are invariant to rotation [5][7]. To achieve scale invariant and translation invariant property, we translate the data points so that the origin is moved to the centroid and scale the points so that the maximum distance from the origin is equal to one [9]. The normalization affects the first two features, $|A_{00}|$ and $|A_{11}|$. From second to fifth order moments, we picked up ten features and they constitute the Zernike feature vector \mathbf{f}_Z :

$$\mathbf{f}_Z = (|A_{20}|, |A_{22}|, |A_{31}|, |A_{33}|, |A_{40}|, |A_{42}|, |A_{44}|, |A_{51}|, |A_{53}|, |A_{55}|). \quad (3)$$

4 Shape Matching

All of the descriptors described above have invariant properties to rotation, translation, and scale changes. Unfortunately the invariant properties of the features hold only for the case of continuous functions, which are not applicable to the digital images. In a discrete case, moment invariants are still invariant under image translation although the moments are computed discretely. But the invariants are expected not to be strictly invariant under rotation and scale changes due to sampling, digitizing, and quantizing of the continuous image for digital computation [7].

There is another critical problem which makes the invariant features not applicable to real images: real images contain some kinds of noises and natural occlusions. From the reason, moment invariants are not sufficient for distinguishing all shapes: they can be very sensitive to noise and their values can be changed drastically with occlusions. Moreover applying Fourier descriptors to images of noises and occlusions is useless in an image retrieval system.

Note that moment invariants are the region-based measures and Fourier descriptors are the boundary-based measures [3]. Mehre et al. [4] showed that the measure using both Fourier descriptors and moment invariants gives the best average retrieval efficiency. They thought this could be because the human perceptual mechanism uses both these aspects of shape in order to compute similarity. We followed the idea of using both types of measures in computing the similarity of two objects.

Contrary to their method of similarity computation, we introduced a two-stage similarity scheme. First, we compute moment invariants to extract relevant images. Then, verification using Fourier descriptors is followed which increases the retrieval effectiveness substantially. The idea of the scheme is that the region-based moment invariants are effective to the classification of appearances and the boundary-based Fourier descriptors are effective to the discrimination of object poses.

The first step computes the moment invariants and excludes images of different appearances. The first n images of the closest distances are selected as candidates for output and other images are excluded. The similarity distance between two feature vectors is computed as the Euclidean distance. The second step computes Fourier descriptors to discriminate object poses. Among the n images, we reorder the sequence using distances of Fourier descriptors and we select $m < n$ images in decreasing order of distances of Fourier descriptors. This process substantially increases the retrieval effectiveness.

5 Experimental Results

To determine the best combination of invariant vectors, we implemented and evaluated all the retrieval effectiveness of the invariant descriptors described in the previous section.

Figure 2 shows the rotation invariant properties and the scaling invariant properties, respectively. The source object is an arrow shaped region in an image

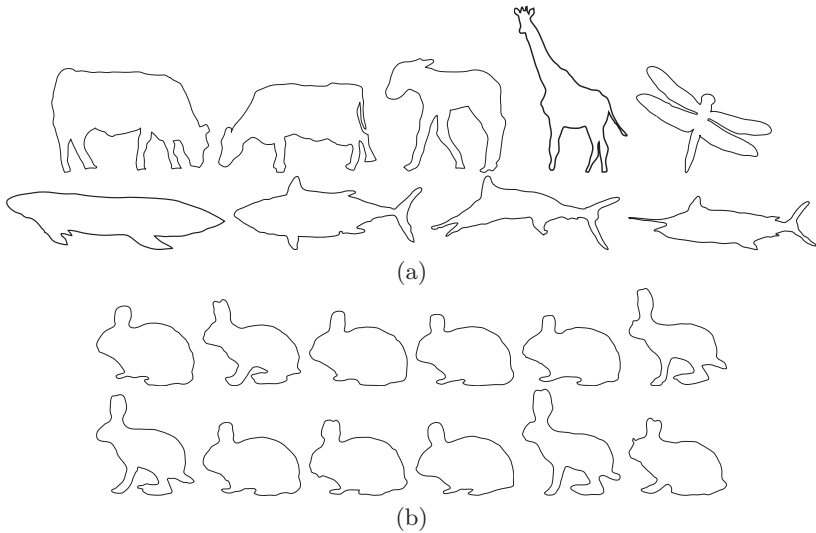


Fig. 3. A test database: (a) nine sample animal images from our experimental database of animal images, (b) twelve target rabbit images to be searched.

of size 337×145 . The transform parameter is a normalized scalar (from 0 to 1) for the translation factor, the rotation factor, and the scaling factor. The translation factor ranges from -400 to 400 in pixel units with step size 8, the rotation factor ranges from 0 to 360 in degree with step size 3.6, and the scaling factor ranges from -2 to 2 with step size 0.04.

All of the descriptors have maintained good invariant properties during the transformation although some moment invariant values oscillated within a narrow range. Although Hu moment invariants \mathbf{f}_H show good invariant properties many features are overlapped which make the features less discriminative. Fourier descriptors \mathbf{f}_F and Zernike moment invariants \mathbf{f}_Z show good discriminating features and few features are duplicated.

As well as two important moment invariants, Hu moment invariants and Zernike moment invariants, described above, we also implemented other moment invariants: Taubin moment invariants and Flusser moment invariants, moment invariants proposed by Taubin and Cooper [10] and those by Flusser and Suk [11], respectively. We used eight dimensional feature space for the Taubin moment invariants [10] and six dimensional feature space for the Flusser moment invariants [11]. We tested many combinations of such features and selected two most promising descriptors:

- $\mathbf{f}_H + \mathbf{f}_F$: Hu moment invariants \mathbf{f}_H plus Fourier descriptors \mathbf{f}_F
- $\mathbf{f}_Z + \mathbf{f}_F$: Zernike moment invariants \mathbf{f}_Z plus Fourier descriptors \mathbf{f}_F

In our experiments, they have given better search results than using a single descriptor or a combination of other descriptors in every cases.

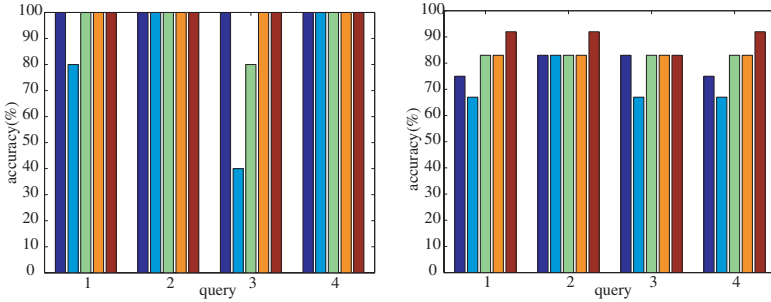


Fig. 4. Comparison of retrieval results. From left to right, the five bars correspond to f_F , f_H , f_Z , f_H+f_F , and f_Z+f_F . (a) Top 5 retrieval results, (b) top 12 retrieval results.

Table 1. Time cost for the computation of the feature vectors. (unit=sec)

feature vector(s)	f_F	f_H	f_Z	f_H+f_F	f_Z+f_F
time cost	0.01	0.02	0.93	0.03	0.94

We tested several different kinds of image databases. Figure 3 shows some of our experimental databases. We constituted a test database which contains about 250 animal images. There are exactly 12 rabbit images in the test database. For the given query rabbit image, statistics of the retrieval accuracy is shown in Fig. 4. Fig. 4(a) shows the number of retrieved rabbit images out of the first five retrieved images and Fig. 4(b) shows the number of retrieved rabbit images out of the first ten retrieved images are shown. In the experiments, the method using both moment invariants and Fourier descriptors showed better results rather than a method using moment invariants only or a method using Fourier descriptors only. The hybrid method f_Z+f_F always gives the best results. Also, another hybrid method f_H+f_F gives at least the second quality results.

We measured the average CPU time for the five similarity measures and the two hybrid measures. The computation time on an Indigo2 IMPACT with a MIPS R10000 processor is shown in Table 1. The test image is of size 337×145 with 718 boundary points and 17,761 region points. Note that the method f_Z+f_F is very slow due to the complex computation of the Zernike moments. Hence, though f_Z+f_F gives the best results, the most practical solution would be f_H+f_F in many applications.

6 Conclusion

In this article we proposed a new hybrid method for the shape-based image retrieval. The method uses both types of Fourier descriptors and moment invariants. We compared the performance with several other popular schemes: Fourier descriptors, Hu moment invariants, and Zernike moment invariants. Experimen-

tal results showed that the proposed hybrid method of Fourier descriptors and Zernike moment invariants is the most effective scheme in the shape-based image retrieval. Another alternative hybrid method of Fourier descriptors and Hu moment invariants is the practical scheme with fast computation and enough search accuracy.

In spite of the strong merits, the proposed method has some drawbacks which could be critical to some sensitive applications. When there is a 3D perspective effect in the shape, unexpected results may appear. The reason of the possible failure is that the invariant properties hold only when the deformation is a kind of 2D affine transformation. Another problem of the method is that the similarity measure heavily depends on the results of a region extraction module. Failure of region extraction makes the shape-based retrieval system give erroneous results.

As future works of our research, we are developing an image segmentation algorithm which extracts only objects of interest regardless of the complexity of the environment where the object is located in. Future research should also include the shape matching of partially recovered objects or objects having occlusions.

References

1. Djeraba, C., Bouet, M., Briand, H., Khenchaf, A.: Visual and textual content based indexing and retrieval. *Int. J. on Digital Libraries* **2** (2000) 269–287
2. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Machine Intell.* **22** (2000) 1349–1380
3. Folkers, A., Samet, H.: Content-based image retrieval using fourier descriptors on a logo database. In Kasturi, R., D.Laurendau, C.Suen, eds.: *ICPR'02*. Volume 3., Quebec City, Canada (2002) 521–524
4. B. M. Mehtre, M.S.K., Lee, W.F.: Shape measures for content based image retrieval: a comparison. *Information Processing and Management* **33** (1997) 319–337
5. N. Ezer, E.A., Sankur, B.: A comparative study of moment invariants and fourier descriptors in planar shape recognition. In: *Proceedings of 7th Mediterranean Electrotechnical Conference*. Volume 1. (1994) 242–245
6. Hu, M.K.: Visual pattern recognition by moment invariants. In: *IRE Transactions in Information Theory*. Volume 8. (1962) 179–187
7. Teh, C.H., Chin, R.T.: On image analysis by the methods of moments. *IEEE Trans. Pattern Anal. Machine Intell.* **10** (1988) 496–513
8. Khotanzad, A., Hong, Y.H.: Invariant image recognition by zernike moments. *IEEE Trans. Pattern Anal. Machine Intell.* **12** (1990) 489–497
9. Chee-Way Chong, P. Raveendran, R.M.: Translation invariants of zernike moments. *Pattern Recognition* **36** (2003) 1765–1773
10. Taubin, G., Cooper, D.B.: Object recognition based on oment (or algebraic) invariants. In Mundy, J.L., A. Zisserman, e., eds.: *Geometric invariance in computer vision*, MIT Press (1992) 375–397
11. Suk, J.F.T.: Pattern recognition by affine moment invariants. *Pattern Recognition* **26** (1993) 167–174

Visual Trigger Templates for Knowledge-Based Indexing

Alejandro Jaimes¹, Qinhui Wang¹, Noriji Kato², Hitoshi Ikeda², and Jun Miyazaki¹

¹ FXPal Japan, Fuji Xerox Co. Ltd, Kanagawa, Japan

{alex.jaimes, noriji.kato, jun.miyazaki}@fujixerox.co.jp

² Corporate Research Laboratory, Fuji Xerox Co., Ltd, Kanagawa, Japan

Abstract. We present an application to create binary *Visual Trigger Templates* (VTT) for automatic video indexing. Our approach is based on the observation that videos captured with fixed cameras have specific structures that depend on world constraints. Our system allows a user to graphically represent such constraints to automatically recognize simple actions or events. VTTs are constructed by manually drawing rectangles to define *trigger spaces*: when elements (e.g., a hand, a face) move inside the trigger spaces defined by the user, actions are recognized. For example, a user can define a raise hand action by drawing two rectangles: one for the face and one for the hand. Our approach uses motion, skin, and face detection algorithms. We present experiments on the PETS-ICVS dataset and on our own dataset to demonstrate that our system constitutes a simple but powerful mechanism for meeting video indexing.

1 Introduction

Meeting videos are important multimedia documents and recently there have been many efforts to set up specialized “smart room” environments to effectively capture meetings [1,2,3,4,5]. One of the major goals of such projects is to automatically index the videos for future viewing. Indexing videos using speech [6] can be extremely useful, but accurate automatic speech recognition in meetings remains a very challenging task due to extensive vocabulary, differences in speech style, topic changes, and difficulties in acquiring high quality audio. Fully indexing videos based on visual content is also a challenging task for which many approaches have been proposed involving face detection, and action recognition, among others.

In a new memory-cue retrieval approach that we are advocating [7], the assumption is that the user of the system will not remember all of the necessary details to find the desired segments in the videos of interest. For example, he may want to review “the part of the meeting when Mr. Biji got special instructions from his boss.” He may not remember at which meeting the instructions were given, but he may remember where people were sitting, that Mr. Biji himself raised his hand to ask a question during a presentation right before the important explanation was given, or that someone used the board right before the

presentation. Querying on these items may yield videos that will give the user further cues that help him remember what he is looking for. In this framework, actions at specific locations are important for indexing (e.g., someone sitting on the corner raised his hand; someone was standing on the left side of the board).

In this paper, we exploit the constraints given by fixed cameras in specialized smart conference rooms with a system that allows a user to graphically construct *Visual Trigger Templates* for visual indexing of meeting videos. Trigger Templates consist of boxes that act as triggers that are used to define specific actions at specific locations (the raising hand action of Fig. 2). When cameras are fixed (stationary and with constant parameters), many actions have a fixed visual structure. A raising hand action captured by a fixed frontal camera, for example, has a structure which varies very little over multiple meetings and persons. Furthermore, the location of many elements in the meeting room remains fixed: where the board is, where the teleconference phone lies, where people sit, etc.

The system we propose is highly flexible: it allows the user to easily create multiple templates for different meeting room scenarios. We apply face, motion, and skin detection algorithms to allow the user to define the templates in terms of boxes and their "on" or "off" trigger conditions. The *on* condition for a raise hand action, for example, consists of a face trigger and a skin trigger near the face.

Our approach differs from previous work in that we exploit the visual structure constraints that result from the capture of videos in specialized meeting rooms. This idea is related to the rule-based expert system of [8]. In our approach, however, rules are constructed with a graphical user interface and the interface is not used for interactive queries [9]. Our focus is on combining object detectors (e.g., face and skin) with constraints provided by the user. A key difference with previous work, therefore, is that we combine two types of knowledge: output of automatic object detectors, and declarative knowledge provided graphically by the user. The authors of [10] apply a similar technique to detect naked people and horses, but focus on skin (and hide) filters and attempt to detect people without exploiting scene structure. Individual and group actions such as stand up and sit down are detected in [11] and [2] using machine learning (for a gesture recognition review see [12]).

2 Visual Trigger Templates

In most of the new smart conference rooms [1,3,4] cameras are placed in fixed locations; the camera locations are chosen carefully so that the meetings are captured well (as little occlusion as possible, good lighting, etc.). Therefore, the visual structure of the images captured by the cameras remains fairly constant (Fig. 1). In [13] this was defined as *Recurrent Visual Semantics*: the repetitive appearance of meaningful visually similar elements (object, events, scenes, etc.) within a specific context.

The user builds rules using a simple drawing tool that allows him to construct Visual Trigger Templates. A *Visual Trigger Template* is a collection of rectangles and "on" conditions for each of the rectangles (we call each rectangle a *trigger*).



Fig. 1. Sample actions and their recurring structure. Note that the structure (face-hand sizes, distances, etc.) is similar for different people sitting at the same location (images on the left for location 1; on the right for location 2)



Fig. 2. Example action recognition training using the interface.

$VTT = \{r_1, \dots, r_n\}$, where $r_i = \{ul, lr, on_condition\}$ where ul , and lr are the (x, y) coordinates of the rectangle’s upper left corner and lower right corner, and $on_condition$ is a feature that must be detected in order for the box to be on (triggered).

Each trigger can be only in one of two states: *on* or *off*. When the user builds the templates he specifies an *on_condition* to satisfy the trigger. An *on_condition* might be, for example, the appearance of an object (e.g., a face) within the trigger area: the trigger is only *on* when a face is detected inside the trigger. In such case, when defining the template, the user just defines the rectangle for the face and presses a “face” button to establish the *on_condition* (since face detection is part of the system). Fig. 2 shows an example of trigger template in on/off states.

Templates can be constructed in terms of *absolute position* or *relative position*. For indexing using particular spatial locations or objects, it is desirable to maintain a fixed absolute position. For example, we know that the projector is at a given location on the table and the projector is fixed. The trigger template that determines when the projector is used can have an absolute location in the video. For a raising hand action, on the other hand, it may be desirable to define the possible area of the raise hand in relation to the face that is detected. In this case the user defines a relative template by graphically drawing it and specifying that the relationship between the trigger rectangles is *relative*.

3 Automatic Processing

The template is fully defined by the user as described above. In applying the templates for automatic indexing a video, the system detects the “trigger” actions corresponding to each of the trigger rectangles.

3.1 Face and Skin Detection

Our face detection system is described in detail in [14,15]. The system consists of an alignment module and a classifier module trained with 1,000 face images. The alignment module normalizes a face candidate with respect to size, position, and angle. The resizing is done within 1 millisecond on a Pentium 4 2.8 GHz machine. Then the normalized face candidate is passed to the classifier. The classifier determines whether the normalized face candidate is a face or not by using a combination of PCA filters (about 15,000 features), which are rich enough for accurate detection. Although the classifier is a “heavy” classifier, our system realizes fast face detection by reducing the number of candidates significantly using the alignment module.

We use the results of the face detector to detect skin areas that are near the faces detected. Once a face is detected, we calculate the histogram of the Hue component of the detected face and detect possible skin pixels using a threshold (i.e., only pixels near the face that are of similar color should be skin pixels).

There are many approaches for face detection, but most face detectors (including our own) yield poor results with non-frontal faces, or with poor lighting (e.g., lights switched off during presentations). Because of this we have also implemented a simple skin detector for detecting limbs and non-frontal faces. Our skin detector is based on the approach described in [13]. We collected several training examples from meeting videos (skin and non-skin areas) and applied a rule-based learning algorithm (using Weka[16]) to partition HSV color space into skin and non-skin areas. Then we used additional training videos and manually improved the rules (e.g., a pixel is skin only if S value is between 0.24 and 0.86 and V value is between 0.5 and 0.98, and so on). Although such customization is not necessary, it improves performance, particularly since the assumption is that the setup of a particular meeting room does not change over time - the type of lighting used is the same, the places where people always sit are similar, and so on. Although we have only tested our skin detector in the PETS and FXPal datasets, similar detectors[17] have been found to be effective in detecting skin independently of the person’s race.

3.2 Feature Extraction

All motion in a meeting room is due to human activity. However, it is not always possible to accurately detect faces and skin areas. In some cases skin may not even be visible (e.g., in detecting someone standing in a particular location and not facing the camera). Therefore, we also detect motion areas using a simple background subtraction algorithm (we use a running time window of length t , find the average between all frames within the window, and subtract the current frame from the average of the window[18]). We group motion and skin pixels (separately) using a run length algorithm and then find the corresponding bounding boxes. We extract basic shape and texture features (e.g., aspect ratio, eccentricity, compactness, etc.[19]). In addition, we use the overlap in time and space between motion and skin bounding boxes to eliminate false skin areas

(humans do not stay completely still for more than a few seconds; see [18]). When detecting skin we eliminate from considerations areas that are too small or below basic shape thresholds (e.g., a minimum compactness). This eliminates many false skin detection areas from consideration and allows us to use motion as a feature in the templates (e.g., trigger “on” if there is motion).

4 Experiments

We tested our approach on the PETS-ICVS dataset[20]and on our own meeting dataset. The PETS dataset contains several videos: for cameras 1 and 2 there are a maximum of 3 people sitting in front of the camera (Fig. 3). Our own database contains one video with several mock-up actions recorded using 5 cameras and involving 4 people.



Fig. 3. Absolute templates built using the system. Notice different people in the same locations.

Using the interface, we constructed several templates (each row in Table 1 corresponds to one template). A different template was constructed *for each camera* and for each action, but not for a particular person. In other words, the “raise right hand template for FX camera 1” was constructed *once* for camera 1. It was then tested on two people who sat on the same chair and performed the same action. The GT (ground truth) column in Table 1 indicates how many actions took place, involving one or more persons. For the FX data set, for each camera, 2 people performed 10 actions each. So the row “Right Hand FX camera 1” includes 20 raise hand actions, 10 by each person. Both persons sat at the same location at different times.

We implemented the templates using 3 different approaches (as indicated in column 1 of Table 1): (1) using a skin filter only (skin raise hand, skin stand up); (2) using the face detector and defining absolute templates (absolute face);

Table 1. Performance of our approach on the Fuji Xerox set (FX) and PETS database (PETS).

Name		GT	Detected	Hit	False Alarms	Miss	Prec. %	Recall %	Acc. %
Absolute Face	Right Hand FX Cam. 1	20	20	20	0	0	100	100	100
	Right Hand FX Cam. 2	20	14	14	0	6	100	70	70
Raise Hand	Right Hand FX Cam. 3	20	20	20	0	0	100	100	100
	Right Hand FX Cam. 4	20	19	19	0	1	100	95	95
	Left Hand FX Cam. 1	20	19	19	0	1	100	95	95
	Left Hand FX Cam. 2	20	20	20	0	0	100	100	100
	Left Hand FX Cam. 3	20	20	20	0	0	100	100	100
	Left Hand FX Cam. 4	20	20	20	0	0	100	100	100
Relative Face	Right Hand FX Cam. 1	20	18	18	0	2	100	90	90
	Right Hand FX Cam. 2	20	14	14	0	6	100	70	70
Raise Hand	Right Hand FX Cam. 3	20	9	9	0	11	100	45	45
	Right Hand FX Cam. 4	20	17	17	0	3	100	85	85
	Left Hand FX Cam. 1	20	18	18	0	2	100	90	90
	Left Hand FX Cam. 2	20	20	20	0	0	100	100	100
	Left Hand FX Cam. 3	20	9	9	0	11	100	45	45
	Left Hand FX Cam. 4	20	1	1	0	19	100	5	5
Skin	PETS Cam. 1	3	3	3	0	0	100	100	100
	PETS Cam. 2	3	3	3	0	0	100	100	100
Stand Up	FX-1	5	4	4	0	1	100	80	80
	FX-2	5	6	5	1	0	83.3	100	83.3
	FX-3	5	5	5	0	0	100	100	100
	FX-4	5	5	5	0	0	100	100	100
	FX-5	5	5	5	0	0	100	100	100
Face	FX-1	5	1	1	0	4	100	20	20
	FX-2	5	2	2	0	3	100	40	40
Stand Up	FX-3	5	5	5	0	0	100	100	100
	FX-4	5	4	4	0	1	100	80	80
	FX-5	5	4	4	0	1	100	80	80
Touch projector (FX)		4	4	4	0	0	100	100	100
Drawing on the board (FX Left)		4	5	3	2	1	60	75	50
Drawing on the board (FX Right)		4	5	4	1	0	80	100	80

(3) using the face detector and defining relative templates (relative face). The drawing actions were detected using only the skin filter to detect the hands (no face detection).

Figure 3 shows example images from the FX data set and the corresponding *absolute* templates.

The errors in the raising hand actions are due to non-detection of the face, or unusual raise hand positions that fall outside the templates. The templates for drawing on the board have a lower performance, but this is most likely because we used only skin detection to define the on_condition to recognize when the participants have their hand raised up against the board. In this case using other features (e.g., amount of motion) would be more beneficial. One drawback of the current implementation is that templates are two-dimensional and rely on simple trigger features. A person walking in front of a board template, for example, would trigger it, even if the template is meant to detect use of the board.

5 Discussion

The trigger templates can be easily constructed using the system's graphical user interface. Once the meeting room is set up, if the cameras are fixed, it

is trivial to construct the templates. Although the approach is very simple as it relies on simple motion, skin, and face detection algorithms, the experiments suggest that it can be a powerful method for indexing meeting videos from smart conference rooms. The method is effective because the layout of the meeting rooms (particularly smart conference rooms) usually remains constant. As the examples show, videos of different people sitting at the same location around a table, at different times, have a very similar visual structure. When relative templates are used, of course, there is more flexibility in the absolute positions of the individuals at the meeting. The relative constraints, however, do not change much.

The templates can also be very powerful for spatially indexing the meeting videos. It is possible to detect, for example, when someone stands at a particular location, or when a particular device in the meeting room is used (e.g., projector). In general, we found the resulting framework to be effective for recognizing actions that involve hand or arm motions, or displacements, and in general whenever we observe fixed structures either due to spatial location or body constraints. Although we do not attempt to learn the templates automatically, it seems like a combination of learning and implicit input could yield interesting results.

6 Conclusions and Future Work

We presented an application to create binary *Visual Trigger Templates* for automatic video indexing. Our approach is based on the observation that images and videos captured with fixed cameras have specific structures that depend on world constraints. The structure of a raise-hand action, for example, depends on our own physical composition: a frontal video shot of a person raising his hand will include a face and a hand within a certain distance to the face. We applied our approach to recognize basic actions in meeting videos (raise hand, stand up, etc.) and presented experiments on the PETS-ICVS dataset[20] and on our own dataset.

Future work includes improving our graphical user interface to deal with template sequences as well as further investigation into the integration of learning algorithms and the definition of temporal trigger templates.

References

1. Liu, Q., Kimber, D., Foote, J., Wylcox, L., Boreczky, J.: Flyspec: A multi-user video camera system with hybrid human and automatic control. In: ACM Multimedia 2002, Juan Les Pines, France (2002)
2. McCowan, I., Bengio, S., D. Gatica-Perez, G.L., Monay, F., D. Moore, P.W., Bourlard, H.: Modeling human interaction in meetings. In: IEEE ICASSP 2003. (2003)
3. A. Waibel, e.a.: Smart: The smart meeting room task at isl. In: IEEE ICASSP 2003. (2003)

4. Mikic, I., Huang, K., Trivedi, M.: Activity monitoring and summarization for an intelligent meeting room. In: IEEE Workshop on Human Motion, Austin, Texas (2000)
5. Jain, R., Kim, P., Li, Z.: Experiential meeting system. In: ACM Multimedia Workshop in Experiential Telepresence (ETP 2003), Berkeley, CA (2003)
6. et. Al., A.W.: Advances in automatic meeting recording and access. In: ICASSP 2001, Salt Lake City, UT (2001)
7. Jaimes, A., Omura, K., Nagamine, T., Hirata, K.: Memory cues for meeting video retrieval. In: CARPE 2004, 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, New York (2004)
8. A. Hakeem, M.S.: Ontology and taxonomy collaborated framework for meeting classification. In: ICPR 2004, Cambridge UK (2004) 23–26
9. S.-F. Chang, W.C., Sundaram, H.: Semantic visual templates: Linking visual features to semantics. In: ICIP 1998, Chicago, Illinois (1998) 4–7
10. Forsyth, D., Fleck, M.: ‘body plans’. In: CVPR-97. (1997) 678–83
11. Zobl, M., Wallhoff, F., Rigoll, G.: Action recognition in meeting scenarios using global motion features. In: IEEE Intl. Wkshp on Perf. Eval. of Tracking and Surveillance (PETS-CCVS), Graz, Austria (2003)
12. Cuzzolin, F.: A gesture recognition review. (<http://www.dei.unipd.it/cuzzolin/Review.html>)
13. Jaimes, A.: Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information. PhD thesis, Department of Electrical Engineering, Columbia University (2003)
14. Ikeda, H., Kato, N., Kashimura, H., Shimizu, M.: Scale, rotation, and translation invariant fast face detection system. In: The 5th IASTED International Conference on Signal and Image Processing. (2003) 146–151
15. Kato, N., Ikeda, H., Kashimura, H.: Tracking faces with rapid movement using appearance based variation detectors. In: Asian Conf. on Computer Vision. (2004) 800–805
16. Witten, I., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2000)
17. Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: Graphicon-2003, Moscow, Russia (2003)
18. Jaimes, A., Yoshida, N., Murai, K., Hirata, K., Miyazaki, J.: Interactive visualization of multi-stream meeting videos based on automatic visual content analysis. In: IEEE Intl. Workshop on Multimedia Signal Processing, Siena, Italy (2004)
19. Russ, J.: The Image Processing Handbook. 3 edn. CRC Press, Boca Raton, Florida (1999)
20. In: 5th IEEE International Workshops on Performance Evaluation of Tracking and Surveillance (PETS-ICVS ICVS’03), Graz, Austria (2003)
<http://petsicvs.visualsurveillance.org/>.

Browsing and Similarity Search of Videos Based on Cluster Extraction from Graphs

Seiji Hotta, Senya Kiyasu, and Sueharu Miyahara

Department of Computer and Information Sciences, Nagasaki University
Bunkyo-machi 1-14, Nagasaki-shi, Nagasaki 852-8521 Japan
{hotta,kiyasu,miyahara}@cis.nagasaki-u.ac.jp

Abstract. This paper presents a browsing and similarity search method of videos based on cluster extraction from graphs. Videos are segmented into shots and represented as the sequence of symbols. The directed graph of videos is constructed from the relationship between those symbols. Initial and terminal shots are extracted from it, and they are displayed for users sequentially from the initial shots to the terminal ones. The shots selected by means of this browsing are used as a query video on similarity search. The performance of the proposed method is examined by using the video dataset of NASA.

1 Introduction

Generally, in content-based video retrieval methods, users input key-frames or query-videos in systems for retrieving desired videos [1]. However, if users do not have such keys, then users continue watching many videos one by one to find desired videos, so temporal-costs for search will be high. Hence, the combination of browsing and search is useful for retrieving videos. The browsing of videos can be divided roughly into two approaches: The first approach outputs sequentially the representative frames of shots to users [2]. The second one outputs all the representative frames of shots to users at the same time [3]. The drawback to the first approach is that we can not know in advance which frames are desired by users. The drawback to the second one is that a great number of frames will be displayed to users when the number of videos is huge.

For overcoming this type of difficulty, we extract introductory shots, major ones and terminal ones for summarization of contents of whole of database videos, and we display their relationship on a low-dimensional space using lines and static images. For this purpose, we use clustering methods that extract fuzzy clusters from graphs. Clustering of graphs is useful for exploration of data structures such as hypertext or similarity between images. In addition, hard clustering is less flexible, since detailed phases among the data are lost, while the condensation efficiency of the data is high. It is believed that fuzzy clustering is especially important for data with high ambiguity such as those related to humans. Hence, fuzzy cluster extraction from graphs is exploited for semantic information extraction and visualization of data structures.

In this paper, first the proposed method segments videos into shots by using fuzzy clustering of undirected graphs and represents videos as the sequence of symbols. The directed graph of videos is constructed from the relationship between those symbols. Initial shots (i.e., the set of shots that have outgoing edges) and terminal shots (i.e., the set of shots that have incoming edges) are extracted from it by fuzzy cluster extraction. Those shots are mapped into a low-dimensional space by applying Principal Component Analysis (PCA) to the membership values of them. For browsing, they are displayed from initial shots to terminal ones. The shots selected by means of this browsing are used as a query video on similarity search. The performance of the proposed method is examined by using the video dataset of NASA.

2 Shot Segmentation by Cluster Extraction

In this section, we will consider shot segmentation of a video using color histograms of each frame. Let the total number of videos in a database and the number of frames of the i th video be V and F_i respectively.

2.1 Quadratic Form Distance Between Histograms

Here, we summarize Quadratic Form Distance (QFD) [4] for measuring the distance between frames. Let the color histogram of the k th frame be $\mathbf{h}_k = [h_{k1}, \dots, h_{kc}]^T$, where c is the number of color bins. We normalize it by $h_{ki} / \sum_{j=1}^c h_{kj}$ for $\sum_{j=1}^c h_{kj} = 1$. The QFD distance between frames k and l is measured by $(\mathbf{h}_k - \mathbf{h}_l)^T \mathbf{C}(\mathbf{h}_k - \mathbf{h}_l)$, where $\mathbf{C} = [c_{ij}]$ is a $c \times c$ matrix and the weight $c_{ij} = \exp(-\alpha D_{LAB}^2(i, j))$ denotes the similarity between color i and j in the CIELAB color space. Throughout this paper, we use the Gaussian function with width parameter $\alpha = 0.001$ for computing this similarity.

2.2 Shot Segmentation by Clustering

For shot segmentation of the i th video, we make use of fuzzy cluster extraction from undirected graphs [5,6]. First, we form a $F_i \times F_i$ matrix defined by $\mathbf{S} = [s_{kl}]; s_{kl} = \exp[-\gamma((\mathbf{h}_k - \mathbf{h}_l)^T \mathbf{C}(\mathbf{h}_k - \mathbf{h}_l) + \beta(t_k - t_l)^2)]$ ($\gamma = 100$). The term $\beta(t_k - t_l)^2$ indicates the temporal dissimilarity between frames k and l , where β and t_k are a scale parameter and the frame number of the k th frame respectively (in experiments, we used $\beta = 10^{-5}$). Next, we calculate the principal eigenvector $\mathbf{u}_1 = [u_{11}, \dots, u_{1F_i}]^T$ of the matrix \mathbf{S} by eigenvalue decomposition. The normalized eigenvector $m_{1i} = u_{1i}/u_{1i_1}$ is used for the membership of the first cluster (i.e., shot) because of its noise robustness [5,6,7], where $i_1 = \arg \max_i \{u_{1i}\}$. That is, the element m_{1i} represents the degree of inclusion of the i th frame in the first shot. In addition, we can use i_1 as the representative frame of the first shot. Note that throughout this paper we use notation ‘‘R-frame’’ as a representative frame for simplification.

In general, the k th shot is extracted by the same procedure after deletion of shots from the first one to the $(k - 1)$ th one; we first calculate the principal eigenvector $\tilde{\mathbf{u}}_k$ of the matrix $\mathbf{S}_k = [s_{ij} \prod_{l=1}^{k-1} \sqrt{(1 - m_{li})(1 - m_{lj})}]$, next $u_{ki} = \tilde{u}_{ki} / \sqrt{\prod_{l=1}^{k-1} (1 - m_{li})}$ is calculated and $i_k = \arg \max_i u_{ki}$ is found, and the membership of the i th frame in the k th cluster is given by $m_{ki} = u_{ki} / u_{ki_k}$. The profile of the variation in the principal eigenvalue of \mathbf{S}_k suggests us an appropriate number of shots (see [5,6] for more details).

Here, we compute the size of shots for similarity search of videos. Let the size of the j th shot be v_j . The value of v_j is calculated by counting the number of frames of which membership value is maximal in the j th shot, so the sum of v_j will be equal to the number of frames F_i .

3 Analysis of Database Videos

3.1 Clustering of R-Frames for Symbol Representation

In this section, we consider symbol representation of each video for extracting initial shots and terminal ones from database videos. Therefore we extract clusters from the set of R-frames. We assign symbols (e.g., number symbols) to extracted clusters and represent videos by sequences of them.

Let the number of extracted shots from the i th video be n_i . Hence, the total number of the extracted R-frames will be $N_f = \sum_{i=1}^V n_i$. Let the color histogram of the p th R-frame be \mathbf{h}_p , and we measure the similarity between R-frames p and q by $s'_{pq} = \exp(-\gamma((\mathbf{h}_p - \mathbf{h}_q)^T \mathbf{C}(\mathbf{h}_p - \mathbf{h}_q)))$ ($\gamma = 100$). We apply the clustering method described in 2.2 to the $N_f \times N_f$ similarity matrix $\mathbf{S}' = [s'_{pq}]$.

After this extraction, each shot in videos is assigned to number symbols based on membership values of R-frames. For instance, when the membership value of the i th R-frame is maximal in the j th cluster, the shot to which that R-frame belongs is assigned to the number symbol j . Note that when the maximal membership of the i th R-frame is smaller than a threshold T ($T = 10^{-5}$), the shot to which that R-frame belongs is ignored in a future processing. In addition, the R-frame of which the membership value is maximal in the j th cluster is referred to as the j th representative shot. Throughout this paper, we use notation ‘‘R-shot’’ as a representative shot.

3.2 Initial Shots and Terminal Shots Extraction by Fuzzy Clustering

Let the total number of symbols be N_s . In order to construct the relationship between symbols in database videos, we form the directed graph by the following manner: When there are K pieces of symbols (denoted by $k'_i (i = 1, \dots, K)$) after the symbol k , we use the inverse number of the number of shots found between k and k'_i as the weight $w_{kk'_i}$. For example, when k'_i is found immediately after k , the weight is computed as $w_{kk'_i} = 1$. On the other hand, when two shots are found between k and k'_i , the weight is computed as $w_{kk'_i} = 1/3$. We calculate weights in

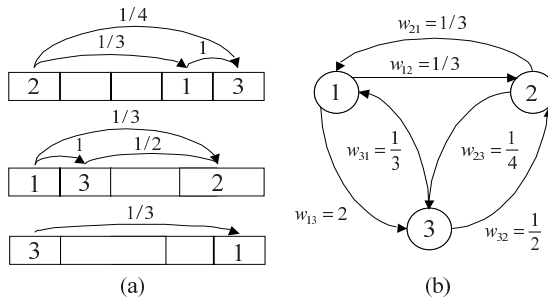


Fig. 1. (a) An example of the relationship of R-shots in database videos. The number of R-shots is three, and blank shots indicate the shots whose memberships is smaller than the threshold T . (b) The directed graph computed based on the relationship between the R-shots depicted in (a).

each video by this manner and form an $N_s \times N_s$ adjacency matrix $\mathbf{W} = [w_{ij}]$ by integrating the weights in each video. An example of this processing is depicted in Fig.1.

In order to extract initial shots and terminal ones from the matrix \mathbf{W} , we use a clustering from directed graphs [6]. The first step of extraction, we calculate $\mathbf{W}\mathbf{W}^T$ and its principal eigenvector $\tilde{\mathbf{p}} = [\tilde{p}_1, \dots, \tilde{p}_{N_s}]^T$. The normalized eigenvector $p_i = \tilde{p}_i / \max\{\tilde{p}_i\}$ is used as the initial membership of the first cluster [6]. On the other hand, the vector $\tilde{\mathbf{q}} = [\tilde{q}_1, \dots, \tilde{q}_{N_s}]^T$ is calculated by $\tilde{\mathbf{q}} = \mathbf{W}^T \tilde{\mathbf{p}}$, and it is normalized by $q_j = \tilde{q}_j / \max\{\tilde{q}_j\}$ for using as the terminal membership of the first cluster [6]. We call $i' = \arg \max_i \{\tilde{p}_i\}$ the representative initial R-shot in the first cluster, which corresponds to the hub in [8], and $j' = \arg \max_j \{\tilde{q}_j\}$ is the representative terminal R-shot which is called the authority in [8]. In general, the k th cluster is extracted by the following procedure: the principal eigenvector $\tilde{\mathbf{x}}_k$ of $W_k W_k^T$ where $W_k = [w_{ij} \prod_{l=1}^{k-1} \sqrt{(1-p_{li})(1-q_{lj})}]$ is calculated and from it we get $\tilde{\mathbf{y}}_k = W_k^T \tilde{\mathbf{x}}_k$ which is further transformed to $y_{kj} = \tilde{y}_{kj} / \sqrt{\prod_{l=1}^{k-1} (1-q_{lj})}$. The initial and terminal membership in the k th cluster are given by $p_{ki} = x_{ki} / \max\{x_{ki}\}$ and $q_{kj} = y_{kj} / \max\{y_{kj}\}$, respectively. The profile of the variation in the square root of the principal eigenvalue of $\mathbf{W}\mathbf{W}^T$ suggests us an appropriate number of clusters (see [6] for more details).

4 Browsing and Similarity Search of Videos

In this section, we show a browsing method that takes advantage of R-shots and its initial and terminal membership. In addition, we show a similarity search that uses the shots selected by this browsing. Let N_g be the number of clusters extracted from a directed graph.

4.1 Browsing

Here, we summarize a browsing method by means of Principal Component Analysis (PCA) for displaying R-shots on a low-dimensional space. Let the combination vector of the initial and terminal membership of the i th R-shot be $\mathbf{x}_i = [p_{1i}, \dots, p_{N_g i}, q_{1i}, \dots, q_{N_g i}]^T (i = 1, \dots, N_s)$. The procedure of PCA for arranging R-shots on a three-dimensional space by this vector \mathbf{x}_i is summarized as follows: First, we compute the average of \mathbf{x}_i by $\mathbf{c} = \sum_{i=1}^{N_s} \mathbf{x}_i / N_s$ and define a $2N_g \times N_s$ matrix \mathbf{X} by $\mathbf{X} \equiv [\mathbf{x}_1 - \mathbf{c}, \mathbf{x}_2 - \mathbf{c}, \dots, \mathbf{x}_{N_s} - \mathbf{c}]$. Next, we calculate the first principal eigenvector \mathbf{e}_1 , the second one \mathbf{e}_2 and the third one \mathbf{e}_3 of $\mathbf{X}\mathbf{X}^T$. Finally, the i th R-shot is displayed on a three-dimensional space by $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]^T (\mathbf{x}_i - \mathbf{c})$. If need be for displaying R-shots from initial shots to terminal ones sequentially, we may rotate the axes of the subspace spanned by \mathbf{e}_1 , \mathbf{e}_2 and \mathbf{e}_3 .

4.2 Similarity Search

Here, we consider a similarity search that uses the R-shots selected via the above browsing method as a query video. In this paper, we measure the distance between the query and videos by Earth Mover's Distance (EMD). EMD is measured by computing the minimal amount of work that must be performed to transform one feature distribution into the other. Let the set of suppliers and that of consumers be I and J respectively. The distance d_{ij} to ship a unit of supply from $i \in I$ to $j \in J$ usually is defined as some kind of distance. Now, we want to seek sets of z_{ij} that minimize the overall cost:

$$\min_{z_{ij}} \sum_{i \in I} \sum_{j \in J} d_{ij} z_{ij}, \quad (1)$$

subject to the following constraints:

$$\sum_{i \in I} z_{ij} = y_j (j \in J), \quad \sum_{j \in J} z_{ij} = x_i (i \in I), \quad z_{ij} \geq 0 (i \in I; j \in J), \quad (2)$$

where x_i is the total supply of supplier i and y_j is the total capacity of consumer j . Let the number of all R-shots selected by users be N_Q . In addition, let v_i^a and v_j^Q be the size of the i th shot in the a th database video and the size of the j th selected R-shot respectively. In this paper, x_i and y_j are $x_i = v_i^a / \sum_{l=1}^{n_a} v_l^a$ and $y_j = v_j^Q / \sum_{l=1}^{N_Q} v_l^Q$ respectively, where n_a is the total number of shots in the a th video. The distance d_{ij} is measured by computing the square root of QFD between the histogram of the i th R-frame \mathbf{h}_i and that of j th one \mathbf{h}_j by $d_{ij} = \sqrt{(\mathbf{h}_i - \mathbf{h}_j)^T \mathbf{C} (\mathbf{h}_i - \mathbf{h}_j)}$. For solving this transportation problem, we use the approximate means proposed by E.J. Russell [9]. For retrieving, we calculate the cost by Eq.(3), and videos sorted in the order of this cost are returned.

$$D_a = \sum_{i=1}^{N_Q} \sum_{j=1}^{n_a} d_{ij} z_{ij} \quad (a = 1, \dots, V), \quad (3)$$

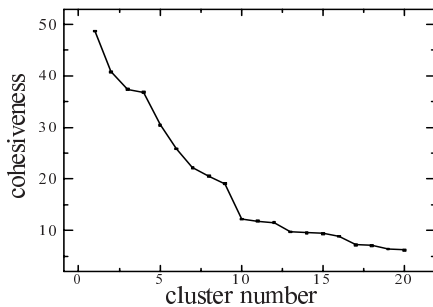


Fig. 2. Variation in eigenvalues of the similarity matrix of frames.

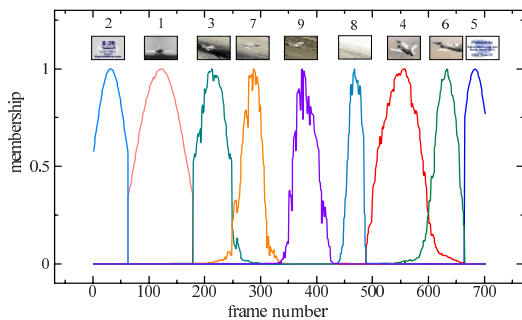


Fig. 3. Shot boundaries.

5 Experiments

We tested the proposed method on videos of aircraft flown at NASA Dryden Flight Research Center (<http://www.dfrc.nasa.gov/Gallery/Movie/index.html>). No copyright protection is asserted for those videos, so we can use them with ease. This dataset contains 166 videos of size 160×120 with 24-bit color resolution, and the average number of frames of them is 932. In experiments, we reduced image resolution and the number of colors to 20×15 and 64 respectively.

5.1 Shot Segmentation of Sample Video

First, each video was segmented into shots by the method described in 2.2. For instance, we show the result on EM-0035-02.mov. Fig.2 shows the variation in the eigenvalue of the similarity matrix of frames. Its decreasing rate deteriorates after the 10th extraction, this shows that clusters from the 10th one are sparse, so the number of salient clusters is nine. Fig.3 shows the membership values of each frame, where attached images and numbers denote the R-frames of each

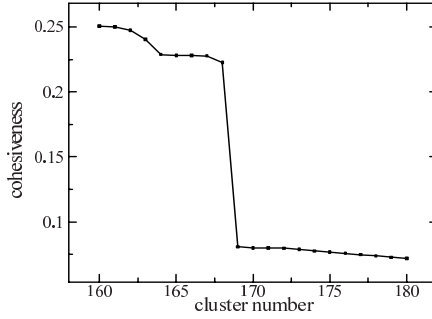


Fig. 4. Variation in eigenvalues of the similarity matrix of R-frames.

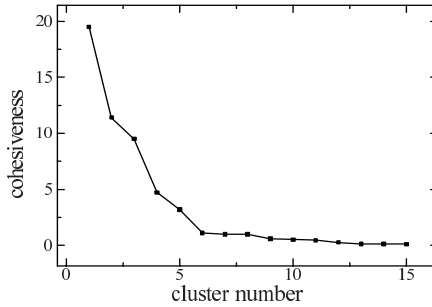


Fig. 5. Variation in eigenvalues of the adjacency matrix of the directed graph.

cluster and the extracted order respectively. Now consider shot boundaries, discontinuous ones mean shot changes such as cuts. On the other hand, continuous boundaries mean successive transition such as fade. By the same manner as this example, all database videos were segmented into shots. Consequently, the total number of extracted R-frames was $N_f = 1588$.

5.2 Cluster Extraction from Sets of R-Frames

Next, we extracted clusters from the $N_f = 1588$ R-frames by the method described in 3.1. Fig.4 shows the variation in eigenvalues of the similarity matrix of R-frames. This figure suggests that the number of clusters (i.e., symbols) is $N_s = 168$.

5.3 Cluster Extraction from Directed Graphs

Next, a directed graph was constructed by the relationship between $N_s = 168$ symbols, and clusters were extracted from it by the method described in 3.2. The variation in eigenvalues of the directed graph is illustrated in Fig.5 from which the number of clusters is determined to $N_g = 5$.

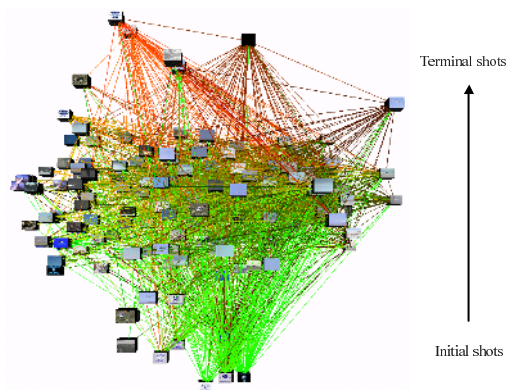


Fig. 6. Arrangement of R-shots by PCA.



Fig. 7. The view from introductory topics.

5.4 Browsing

Next, we applied the browsing technique described in 4.2 to this dataset. Fig.6 shows the arrangement of R-shots and their directed edges. The R-shots of which edges in green have links to many other R-shots, so they serve to introduction to topics of database videos. On the other hand, the R-shots which have edges in red are linked from others, so their contents provide users with terminal topics of database videos. In addition, the R-shots which have edges in yellow have links to others and are also linked from others. Hence, their contents provide users major topics in database videos.

Fig.7 shows the view from the initial point of the arrow in Fig.6. In other words, this figure is the view at which it is looked from introductory topics. In



Fig. 8. Retrieval result.



Fig. 9. Retrieval result.

this area, a lot of logos of NASA which appear at start points of video sequences are shown, so we can conclude that this arrangement is an adequate result.

5.5 Similarity Search of Videos

In this section, we show the result of similarity search of videos by the method described in 4.2. Let us begin with a simple case with a query which includes the scene of “take-off”. For this case, users will select some images by the above browsing, which include a land surface and an aerial region. The top rows in Fig.8 and Fig.9 show the example of R-shots selected by a user. Those R-shots include a land surface and an aerial region, and we retrieved videos by using them. The retrieved videos are shown in Fig.8 by their R-frames in time-series order. This result is unsatisfactory because not the scene of “take-off” but that of “landing” is contained. Hence, for improving the retrieval performance, we changed the supply of suppliers and the total capacity of consumers using the initial and terminal membership by the following manner: Let the symbol number of the i th frame of the a th video be k_i . We use notation $p_{k_i}^a$ as the maximal initial membership value of the symbol k_i . On the other hand, let the symbol number of the j th selected shot of a query be k_j . We use notation $p_{k_j}^Q$ as the

maximal initial membership value of the j th selected R-shot. First, we weighted v_i^a and v_j^Q by using $p_{k_i}^a$ and $p_{k_j}^Q$ as follows; $v_i^a \leftarrow v_i^a p_{k_i}^a$ and $v_j^Q \leftarrow v_j^Q p_{k_j}^Q$. After the weighting, we compute the cost defined by Eq.(3) (denoted by D_a^I). Next, we weighted the v_i^a and v_j^Q by the maximal terminal membership values and compute the cost in the same manner as the initial ones. This cost is denoted by D_a^T . Finally, we combine those costs as $D_a^I + D_a^T$, and systems return the videos sorted in the order of this value. This combined cost will be small when the color histograms and the distribution of the initial and terminal membership in a query and database videos are similar. Fig.9 shows the retrieval result obtained by this modification. This result is satisfactory one because those videos include the scene of “take-off” only. Hence, we can conclude the initial and terminal membership helped improve the retrieval performance.

6 Conclusions

This paper has presented a browsing and similarity search method of videos based on cluster extraction from graphs. It was verified by experiments that the initial and terminal membership of fuzzy clusters were effective for browsing and similarity search. However, we did not explore both about an evaluation measure for the proposed method and enough comparison to the other similar method. The future work will be dedicated to finding an effective evaluation measure for the proposed method.

Acknowledgments. This research was supported by The Ministry of Education, Culture, Sports, Science and Technology in Japan under a Grant-in-Aid for Scientific Research No.15700102.

References

1. Dimitrova, N., Abdel-Mottaled, M.: Content-based video retrieval by example video clip. Proc. SPIE **3022** (1997) 59–70
2. Zhang, H. J., Wu, J., Zhong, D., Smolier, S.: An integrated system for content-based video retrieval and browsing. Pattern Recog. **30**(4) (1997) 643–658
3. Lin, T., Zhang, H. J., Shi, Q. Y.: Video content representation for shot retrieval and scene extraction. Proc. ICIP-94, **2** (1994) 66–70
4. Sawney, H. S., Hafner, J. L.: Efficient color histogram indexing. Proc. ICIP-94 **2** (1994) 66–70
5. Inoue, K., Urahama, K.: Sequential fuzzy cluster extraction by a graph spectral method. Patt.Recog Lett. **20**(7) (1999) 699–705
6. Hotta, S. Inoue, K., Urahama, K.: Extraction of fuzzy clusters from weighted graphs. Proc. PAKDD-2000 Terano T. (Ed.) (2000) 442–453
7. Tsuda, K., Minoh, M.: Extracting straight lines by sequential fuzzy clustering. Patt.Recog Lett. **17** (1996) 643–649
8. Kleinberg, J. M.: Authoritative sources in a hyperlinked environment. Proc. SODA’98 (1998) 668–677
9. Russell, E. J.: Extension of Dantzig’s algorithm to finding an initial near-optimal basis for the transportation problem. Operations Res. **17** (1969) 187–191

Correlation Learning Method Based on Image Internal Semantic Model for CBIR

Lijuan Duan¹, Guojun Mao¹, and Wen Gao^{2,3}

¹ The College of Computer Science, Beijing University of Technology,
Beijing 100022, China

{ljduan,maoguojun}@bjut.edu.cn

² Institute of Computing Technology, Chinese Academy of Sciences,
Beijing 100080, China

wgao@ict.ac.cn

³ Graduate School, Chinese Academy of Sciences, Beijing 100039, China

Abstract. Semantic-based image retrieval is the desired target of Content-based image retrieval (CBIR). In this paper, we proposed a new method to extract semantic information for CBIR using the relevance feedback results. Firstly it is assumed that positive and negative examples in relevant feedback are containing semantic content added by users. Then image internal semantic model (IISM) is proposed to represent comprehensive pair-wise correlation information for images through analyzing the feedback results. Finally, correlation learning method is proposed to represent the images' pair-wise relationship based on statistical value of access path, access frequency, similarity factor and correlation factor. Experimental results on Corel datasets show the effectiveness of the proposed model and method.

1 Introduction

In the past few years, Content-Based Image Retrieval (CBIR) has been a very active research area [1-7]. Semantic feature extraction and description is the headstone of CBIR. The semantic feature extraction methods can be subdivided into: automatic or semi-automatic image segmentation and understanding, automatic or semi-automatic keyword annotation. The classical systems of former include SIMPLIcity [1], SceneryAnalyzer [2] and Blobworld [3]. Although these systems [1-3] support semantic retrieval, the application fields are limited. They can only extract simple semantic features, such as types semantic or object. By now, there is no a universal object recognition method; image understanding is still an open problem. For human, it is not difficult to extract the semantic information from an image. The background knowledge plays an important role in human object recognition. So, manual or semi-automatic image annotation methods are adopted by some CBIR systems [4-6]. But the semantic meaning of an image is very rich. So, it is said that an image is worthy of thousand words. Sometimes, an image is given different meaning by different people. Even one people can give an image different semantic meaning in different time or different situation.

Relevance feedback is an excellent technique for improving the retrieval effectiveness. But most feedback approaches do not have a learning mechanism to memorize the feedbacks conducted previously and reuse them in favor of future queries [8]. A lot of relevance feedback retrieval systems need many feedbacks to obtain satisfied results. In fact, a lot of people dislike giving so many times feedbacks. If we can record the historic relevance feedbacks information and analyze them, the performance of CBIR will be enhanced. We propose a novel semantic model, image internal semantic model (IISM). IISM extracts the semantic information not by image segmentation and image understanding, but by analyzing relevance feedback image retrieval results. For relevance feedback image retrieval system, the images relevant to query are pointed as positive example, otherwise the images irrelevant to query are pointed as negative examples. It is assumed that these positive examples are related in semantic content. IISM provides semantic based image retrieval by collecting and analyzing the relevance feedback retrieval results.

How to correlate relevance feedback information together and how to extract user's common access action is the main problem for IISM. The noise information must be eliminated. In this paper, we introduce a correlation learning method to support IISM. It represents images' pair-wise relationship based on statistical value of access path, access frequency, similarity factor and correlation factor. Then, user can directly access image's correlation information to implement semantic retrieval.

The rest of this paper is organized as follows. Section 2 gives the definition of IISM. Section 3 presents our algorithm to extract semantic correlation based on log sequence. The effectiveness of our algorithm is presented in Section 4. Section 5 concludes the paper.

2 Image Internal Semantic Model

The basic idea of IISM is extracting the relationships between images. For relevance feedback image retrieval system, the images relevant to query are pointed as positive example, otherwise the images irrelevant to query are pointed as negative examples. It is assumed that these positive examples are related in semantic content. IISM computes comprehensive pair-wise correlation information for images through analyzing the results of relevance feedback image retrieval. An association with high value means that one image is semantically associated with another. The following definition displays the up mentioned idea formally.

2.1 Image Retrieval Transaction

In information system, transaction is a useful concept, which is the basic logical operational unit. Image retrieval transaction is a complete query process for relevance feedback image retrieval system.

The outcome R_i of relevance feedback image retrieval is represented by an image retrieval transaction $\langle Q, P, N, I, S \rangle$. Q is query example.

P is positive examples, $P = \{p_1, p_2, p_3, p_4, \dots, p_n\}$. N is negative examples, $N = \{n_1, n_2, n_3, n_4, \dots, n_l\}$. I is the retrieved images $I = \{i_1, i_2, i_3, i_4, \dots, i_m\}$. $S = \{s_1, s_2, s_3, s_4, \dots, s_m\}$ is the similarity value between Q and each retrieved images in I .

2.2 Semantic Correlation Images

The images related in semantic level are called semantic correlation images. Such as Q and $P = \{p_1, p_2, p_3, p_4, \dots, p_n\}$ are semantic correlation images. Q and $N = \{n_1, n_2, n_3, n_4, \dots, n_l\}$ are not semantic correlation images.

For a relevance feedback image retrieval system, it is not reality to capture the semantic relationship of all images by analyzing one query or one feedback. Even for the same query, different user has respective interest, so the query result may not be consistent with others. The “correct” semantic of a multimedia is what most people (but not necessarily all the people) agree upon [7]. Noisy information must be identified and eliminated. In next section, a correlation learning method is introduced to extract semantic correlated images by analyzing old-time query logs.

3 Correlation Learning Method

Based on IISM Image internal semantic model assumes that positive examples are related in semantic content. So, we can compute comprehensive pair-wise correlation information for images through analyzing the feedback results. How to correlate relevance feedback information together and how to extract user’s common access action is the main problem for IISM.

In this section, correlation learning method is proposed to represent the images’ pair-wise relationship based on statistical value of access path, access frequency, similarity factor, and correlation factor. Path represents the relationship between query image and other images. Access frequency evaluates the user’s feedback frequency about a path. Similarity factor represents two images’ similarity degree. Correlation factor synthesizes the access frequency and similarity factor, which is the base to rank images in semantic learning process.

3.1 Path and Access Frequency

Path and access frequency is often used in information processing. Here, path represents the relationship between query image and other images. Access frequency evaluates the user’s feedback frequency about a path.

i_q is the query example. If the retrieval system finds that i_j is similar to query example, then there is a path from i_q to i_j . $path(i_q, i_j)$ represents the path from i_q to i_j . As we known, not every path is interesting. Usually, user will give feedback after each query process. It is very natural that different feedback processes induce different retrieval results even for the same query example. Some

paths are often visited, while some paths are visited seldom. Access frequency is used to distinguish the approbatory degree of each path. It is defined as follows.

$$F(i_q, i_j, t) = \frac{N_{qj}^+ - N_{qj}^-}{\max_{k \in [1, m]} (N_{qk}^+ - N_{qk}^-)} \quad (1)$$

$F(i_q, i_j, t)$ represents the access frequency of $path(i_q, i_j)$ from original time to time t , which can be used to evaluate the interesting degree of a path. N_{qj}^+ represents the feedback times that users regard image i_q is similar to i_j , which is called positive feedback times. N_{qj}^- represents the feedback time that users regard image i_q is dissimilar to i_j , which is called negative feedback times.

In equation 1, if the denominator is equal to zero, then $F(i_q, i_j, t) = 0$. Usually, the value of access frequency between image i_q and i_j is normalized to $[-1, +1]$. If the positive feedback time is less than negative feedback times, then the value of access frequency between image i_q and i_j is negative. It is illustrated that the retrieval system regard that i_q is similar to i_j , but user regard that they are dissimilar.

3.2 Similarity Factor

Similarity factor is another element to evaluate each path, which represents similarity of each path vertexes. If $path(i_q, i_j)$ and $path(i_q, i_p)$ has same access frequency, but sometimes $similarity(i_q, i_j)$ and $similarity(i_q, i_p)$ is different. Different RF algorithm will induce different image rank order and different similarity values. Even for the same RF algorithm, different feedback process will induce different similarity values. How to conform the similarity information of each retrieval process is very important. In order to eliminate the noise information and grasp main information, a global similarity factor is defined as follows.

At time t , $L(i_q, i_j, t)$ is defined as follows,

$$L(i_q, i_j, t) = W_1 * L(i_q, i_j, t - 1) + W_2 * S(i_q, i_j, t) \quad (2)$$

$L(i_q, i_j, t - 1)$ represents the evaluation for $path(i_q, i_j)$ at time $t - 1$. $S(i_q, i_j, t)$ is the similarity between i_q and i_j at time t . W_1 is the weight of $L(i_q, i_j, t - 1)$. W_2 is the weight of $S(i_q, i_j, t)$. The affection degree of historic information can be decided by modified W_1 and W_2 . Global similarity factor can not only reflect the evaluation from all users for the path, but also embed time concept to fade the older knowledge from memory.

3.3 Correlation Factor

If there are enough users that had accessed enough images, the access path and similarity factor becomes clear and clear. By analysis these logs, we can find out which images are often accessed, and which images are similar to other images. In another word, the historic retrieval information can instruct other user's image retrieval. Correlation factor integrated the two elements: access

frequency and similarity factor. It is the bases to ranking the images in semantic learning process. It is the weighted sum of access frequency and similarity factor. The formula is as follows.

$$C(i_q, i_j, t) = W_3 * L(i_q, i_j, t) + W_4 * F(i_q, i_j, t) \quad (3)$$

W_3 is the weight of $L(i_q, i_j, t)$. W_4 is the weight of $F(i_q, i_j, t)$.

4 Experiment Result

In order to test the efficiency of mentioned algorithm, an experiment system is established, which adopts the Rich Get Richer (RGR) relevance feedback strategy [9]. Tests are performed on commercial database Corel Gallery. It contains 1,000,000 images, being classified into many semantic groups. We create a test database by randomly selecting 20 categories of Corel Photographs (30 images in each category). 7196 times retrievals are collected.

Access path and access frequency are extracted based on the 7196 retrieval transactions. For each path, similarity factor and correlation factor is computed by using the method mentioned in section 3. Then, semantic retrieval is implemented. 300 images are selected as query images. For each of these 300 query images, the first 30 images are examined. Table 1 shows the semantic retrieval results by selected different parameters. It displays that more than 77% images correlated to query image can be retrieved by using correlation learning method.

Table 1. Semantic retrieval result by using correlation learning method

W_1	W_2	W_3	W_4	Average Precision
0.3	0.7	0.3	0.7	76.3%
0.3	0.7	0.4	0.6	78.9%
0.3	0.7	0.5	0.5	78.2%

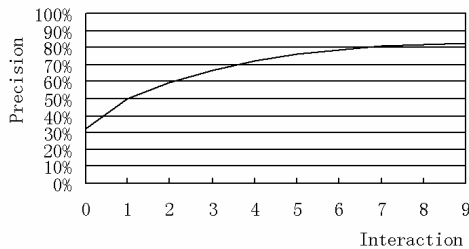


Fig. 1. Average performance of the image retrieval by using RGR. The y-axis represents the average precision over 300 retrievals. The x-axis represents the number of interaction.

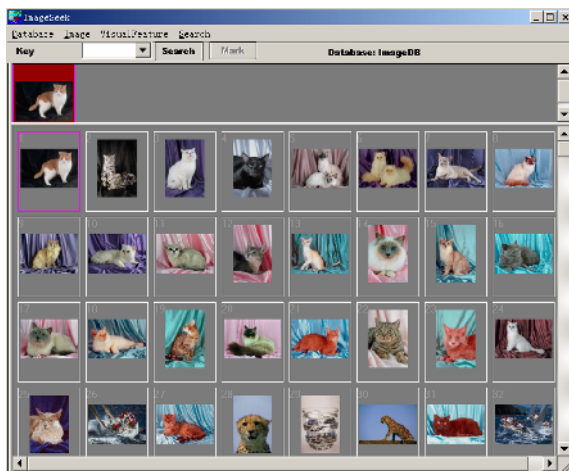


Fig. 2. Retrieval result by using correlation learning method

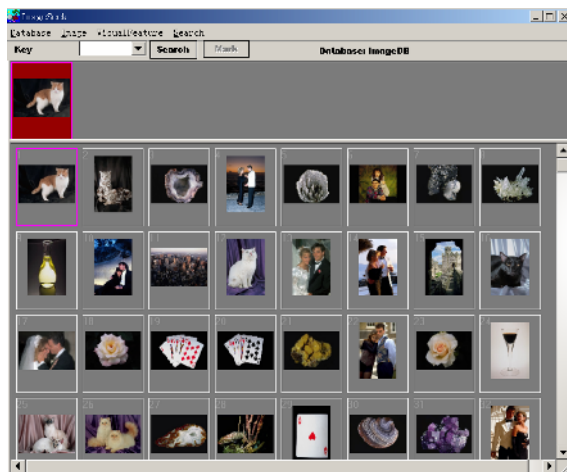


Fig. 3. Retrieval result by using RGR based on visual feature

To illustrate the efficiency of correlation learning method based on IISM. A comparison experiment is done on the same image database. It executes image retrieval by using the RGR [9] method. As shown in Fig. 1, the average precision based on visual feature vectors is 32%, while if the system incorporates relevance feedback from the user, the precision is up to 75% after 5 interactions. It is obviously that the semantic correlation learning method based on IISM can improve the retrieval efficiency and enhance users' satisfaction.

The detail of query process is given in Fig.2 and Fig.3. Fig. 2 gives an example of semantic retrieval using correlation learning method mentioned in section 3.

Fig. 3 gives an example of visual feature based retrieval by using RGR [9]. In Fig.2 and Fig. 3, the query image is displayed at the upper left corner. As these show that the query images of the two experiments are same, which comes from the “cat” category of Corel database. The best 30 retrieved images are displayed. In Fig.2, 90% images are belonging to “cat” category. While in Fig.3, 80% images are not belonging to the “cat” category. The performance of the query by using RGR can up to 90% after 10 interactions (In this section, the details of feedback query process displaying is omitted). In fact, most people dislike so many times feedback. It obviously shows that the correlation learning method is very useful for semantic retrieval.

5 Conclusion

In this paper, a correlation learning method based on Image Internal Semantic Model is presented for semantic image retrieval.

IISM extracts the semantic information not by image segmentation and image understanding, but by analyzing relevance feedback image retrieval results. For relevance feedback image retrieval system, the images relevant to query are pointed as positive example, otherwise the images irrelevant to query are pointed as negative examples. It is assumed that these positive examples are related in semantic content. The correlation learning method computes images’ pair-wise association information through analyzing the log sequence of relevance feedback image retrieval. It uses access path, access frequency, similarity factor and correlation factor to represent images’ pair-wise relationship. A correlation with a high degree means that one image is semantically associated with another closely. Semantic retrieval is carried out based on these correlation relationships. Primary experiments on the database with 600 images show that it is efficient and simple. In the future, we will test it on larger database.

Acknowledgment. This paper is supported by the Ph. D. Research Fund and Youth Research Fund of Beijing University of Technology.

References

1. Wang, J. Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive Integrated Matching for Picture Libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 23 (Sept. 2001) 947–963
2. Song Y., Zhang A.: SceneryAnalyzer: a System Supporting Semantics-based Image Retrieval. In *Intelligent Multimedia Documents*. C. Djeraba, Kluwer Academic Publishers (2002) 43-58
3. Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., Malik, J.: Blobworld: A system for Region-based Image Indexing and Retrieval. *Third International Conference on Visual Information Systems* (June 1999) 509–516
4. Cox, I. J., Miller, M. L., Omohundro, S. M., Yianilos, P. N.: Pichunter: Bayesian Relevance Feedback for Image Retrieval System. In *International Conference on Pattern Recognition*. Vienna, Austria (Aug. 1996) 361-369

5. Lu, Y., Hu, C.-H., Zhu, X.-Q., Zhang, H.-J., Yang, Q.: A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems. *ACM Multimedia* (2000) 31-37
6. Yang, J., Liu, W.-Y., Zhang, H.-J., Zhuang, Y.-T.: An Approach to Semantics-Based Image Retrieval and Browsing. *International Workshop on Multimedia Database Systems*. Taipei, Taiwan (Sept. 2001)
7. Li, Q., Yang, J., Zhuang, Y.-T.: Web-Based Multimedia Retrieval: Balancing Out between Common Knowledge and Personalized Views. *Proceedings of the 2nd International Conference on Web Information Systems Engineering (WISE'01)*. Kyoto, Japan, Vol. 1 (Dec. 2001) 92-101
8. Zhuang, Y.-T., Yang, J., Li Q., Pan, Y.-H.: A Graphic-Theoretic Model for Incremental Relevance Feedback in Image Retrieval. *Proceeding of 2002 International Conference on Image Processing (ICIP 2002)*. New York (Sep. 2002) 22-25
9. Duan, L.-J., Gao W., Ma, J.-Y.: A Rich Get Richer Strategy for Content-Based Image Retrieval. *Fourth International Conference on Visual Information Systems*. Lyon, France (Nov. 2000) 290-299

Controlling Concurrent Accesses in Multimedia Databases for Decision Support

Woochun Jun¹ and Suk-ki Hong²

¹ Dept. of Computer Education, Seoul National University of Education, Seoul, Korea
wocjun@ns.snue.ac.kr

² Division of Business & Economics, Dankook University, Seoul, Korea
skhong017@dankook.ac.kr

Abstract. The decision support processing is essential in multimedia databases since it reveals valuable information from tremendous hidden data. In decision support environments, most transactions have long-term read operations accessing significant portions of database. In this sense, the traditional concurrency control schemes that are tuned to online transaction processing (OLTP) are not suitable for decision supporting environments since long transactions may cause serious locking overhead. In this paper, a locking-based concurrency control scheme is presented for decision support environments in multimedia databases. In this work, transactions are classified into two groups, the typical OLTP transaction and query transaction that is composed of read operation for decision support. Assuming that query transactions read considerable portions of whole database, the proposed scheme incurs less locking overhead than the existing scheme called explicit locking. This paper also proves that the proposed scheme performs better than the existing scheme.

1 Introduction

Multimedia databases provide the database capabilities of accessing, concurrent sharing of multimedia information. When we consider multimedia datatypes, there is a natural association of object-oriented concepts with multimedia as follows [7]. At first, object-oriented databases (OODBs) attempt to model the real world as closely as possible. This is also the goal of multimedia applications. Second, multimedia datatypes have a specific set of operations that are applicable to each datatype. The association of operations with a “type” is a fundamental paradigm in object-orientation. The third reason is due to organizational aspect of object-orientation and multimedia. That is, objects are organized in various collections. The multimedia object can be a subpart of many graph-structured object spaces.

In decision support environments, most transactions (we call query transactions) are to extract the snapshots of operational data. Those query transactions are usually long-running and read significant portions of database [3,5]. Thus, with the traditional locking-based concurrency control schemes, processing such a long-running transaction may cause the serious locking overhead. In turn, this overhead may cause the serious performance degradation. In this sense, it is necessary to develop a concurrency control

scheme incurring less locking overhead. In this paper, we present a locking-based concurrency control scheme for decision support environments in multimedia databases. Especially, our scheme deals with OODBs.

One of the major characteristics of OODBs is inheritance hierarchy. That is, a subclass inherits definitions defined on its superclasses. Also, there is an *is-a* relationship between a subclass and its superclasses so that an instance of a superclass is a generalization of its subclasses [4]. This inheritance relationship between classes forms a class hierarchy. In this paper, for our concurrency control scheme, we assume that there are two kinds of transactions in OODBs, the OLTP transaction and the query transaction. The OLTP transaction is a traditional transaction that reads and writes a few data items of database. On the other hand, the query transaction consists of a set of successive read-only queries [8]. Due to inheritance hierarchy, for a lock-based concurrency control scheme, when a query transaction is requested on some class C, it may be necessary to get locks for C as well as all subclasses of C.

In the literature, there are two approaches dealing with class hierarchy in OODBs, explicit locking and implicit locking, which will be discussed in Section 2. These approaches may work well only for particular applications in OODBs. That is, explicit locking incurs less locking overhead for transactions invoking mostly OLTP transactions. On the other hand, implicit locking incurs less locking overhead for query transactions. In this paper, we present a new concurrency control scheme and prove that the proposed scheme incurs less locking overhead than explicit locking.

This paper is organized as follows. In Section 2 we review previous works dealing with class hierarchy. In Section 3 a new concurrency control scheme is proposed. We also prove the correctness of our scheme. In Section 4, it is shown that the proposed scheme performs better than existing scheme. The paper concludes with future research issues in Section 5.

2 Related Works

In the literature, there are two major locking-based approaches dealing with a class hierarchy: explicit locking [2,10] and implicit locking [4,6,9]. In explicit locking, for a query transaction on a class, C, a lock is set not only on the class C, but also on each subclass of C in the class hierarchy. For an OLTP transaction, a lock is set for only the class to be accessed (called target class). Thus, query transactions accessing a class near the leaf in a class hierarchy will require fewer locks than transactions accessing a class near the root in the class hierarchy. Also, another advantage is that it can treat single inheritance and multiple inheritances in the same way. But, explicit locking incurs more locking overhead for transactions accessing a class near the root in a class hierarchy.

On the other hand, the implicit locking is based on intention locks [1]. The purpose of an intention lock on a class indicates that some lock is set on a subclass of the class. Thus, when a lock is set on a class C, it is required to set extra locking on every superclass of C as well as on C. In implicit locking, when a query transaction need to access class C, locks are required on the class C (in single inheritance) or locks on C and its subclasses having more than one superclass (in multiple inheritance) [4]. Thus, for a query transaction, it incurs less locking overhead than explicit locking. But, due

to intention locking, implicit locking incurs more locking overhead when a target class is near the leaf in a class hierarchy.

3 Proposed Class Hierarchy Locking Scheme

3.1 Background

The proposed scheme is based on the explicit locking scheme. The reason we do not consider the implicit locking is as follows. There have been some works to reduce the intention locks in implicit locking [6]. While those schemes reduce locking overhead, they may affect the original serialization order. In this work, the basic idea is that some redundant locks can be reduced without affecting the correctness of the scheme and also serialization order.

Assume that a class C is accessed so that it needs to be locked. For the explicit locking scheme, depending on the OLTP or query transaction, locks are required for each subclass of C as well as the class C itself. On the other hand, the proposed scheme does not have to set locks on every subclass of C . That is, only some designated classes (denoted as DC) are locked.

3.2 A New Class Hierarchy Locking Scheme

Based on the ideas explained as in Section 3.1, the proposed scheme is as follows. Assume that a lock is requested on class C . Also, it is assumed that the strict two-phase locking scheme is adopted [1].

Step 1) Locking on a target class

- Check conflicts with locks set by other transactions and set a lock on the target class. Also, set a lock on an instance if necessary.

Step 2) Locking on DCs

- Lock requester is an OLTP transaction

For the first DC through subclass chain of C , check conflicts with locks set by other transactions and set a lock on the class and set a lock on the instance if necessary.

- Lock requester is a query transaction

- For each class of C from subclass of C to the first DC through subclass chain of C , check conflicts with locks set by other transactions and set a lock on the class and set a lock on the instance if necessary.

- For each DC through subclass chain of C , check conflicts with locks set by other transactions and set a lock on the instance if necessary.

For example, consider the class hierarchy as in Fig. 1.a. Also, assume that locks are requested by T_1 and T_2 as follows.

T_1 : OLTP transaction on class $C5$, T_2 : query transaction on class $C1$

As in Fig 1.a and Fig. 1.b, 7 and 10 locks are required for T_1 and T_2 by the proposed scheme and explicit locking, respectively. Assume that L_i be a lock L set by transaction T_i .

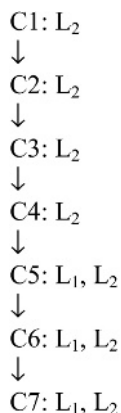


Fig. 1.a. Locks by the explicit locking

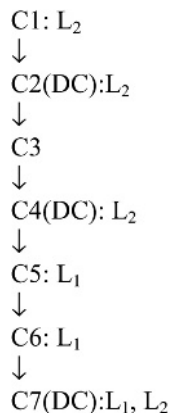


Fig. 1.b. Locks by the proposed Scheme

3.3 DC Assignment Scheme

In this section, we present a scheme to determine whether a class can be DC or not. Assuming that the number of accesses to each class is stable and the access frequency of the OLTP transaction and query transaction to each class is known in advance, the DC assignment scheme is explained as follows.

//Start from each leaf in the class hierarchy until all classes are checked//

Step 1) If a class is a leaf, then the class is assigned as DC.

Step 2) If a class is a root, then the class is assigned as non-DC.

Otherwise, do the following.

If a class C has not been assigned yet and all subclasses of C have been already assigned, then do the following:

for class C and all of its subclasses,

calculate the number of locks (N_D) when the class is assigned as DC

calculate the number of locks (N_N) when the class is assigned as non-DC

Step 3) Assign it as DC only if $N_D < N_N$. Repeat Step 2) until all classes are checked.

For example, consider a simple single-inheritance class hierarchy as in Fig. 2.a and assume access frequency information on each class is known as in Fig. 2.b. The DC assignment to each class is as follows. First C1 is assigned as DC since C1 is a leaf class. Consider the class C2. If C2 is assigned as non-DC, the numbers of locks needed for class C1 and C2 are 200 (for C1) and 500 (for C2), respectively, resulting 700 locks. On the other hand, if C2 is assigned as DC, then locks needed for classes C1 and C2 are 800 locks, where 200 locks are for C1 (100 locks for query transaction and 100 locks for OLTP transaction) and 600 locks are for C2 (200 locks for OLTP transaction and 400 locks for query transaction). Thus, C2 becomes non-DC. Similarly, assigning class C3 as DC and non-DC incur 2,200 locks and 2,900 locks, respectively. Thus, C3 is assigned as DC. Fig. 2.c shows the result of the DC assignment scheme.



Fig. 2.a. A class hierarchy

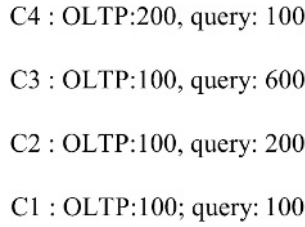


Fig. 2.b. Access frequency for each class

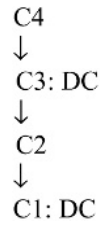


Fig. 2.c. Results of DC assignment

3.4 The Proof of Correctness of the Proposed Scheme

In this section, we show that our scheme is correct. That is, we show that our scheme satisfies the serializability. Since our scheme is based on strict two-phase locking, it is sufficient to prove that, for any lock requester, the possible conflict with a lock holder is always detected.

Claim) For any lock requester, the possible conflict with a lock holder is always detected by the proposed scheme.

Proof) There are four cases depending on the transaction types of lock requester and lock holder

Case 1) Both are OLTP transactions

In this case, if both transactions have the same target class, the possible conflicts are always detected on the target class. Otherwise, there is no conflict between a lock holder and a lock requester.

Case 2) Lock requester: OLTP transaction, lock holder: query transaction

In this case, if the target class of the lock requester is superclass of the target class of the lock holder, there is no conflict. Otherwise, there are two cases as follows.

Subcase 2.1) There is no DC between the lock holder and lock requester

In this case, the possible conflict is detected on the target class of lock requester. For example, in Fig 3.a, let C6 and C5 denote the target class of a lock requester T_R and a lock holder T_H , respectively. In fig. 3.a, the conflict is detected on the class C6.

Subcase 2.2) There is at least one DC between the lock holder and lock requester

In this case, the possible conflict is detected on the first DC of the lock requester through subclass chain of the lock requester. Let C3 and C1 denote the target class of lock requester T_R and a lock holder T_H , respectively. In Fig. 3.b, the conflict is detected on the class C4.

Case 3) Lock requester: query transaction, lock holder: OLTP transaction

In this case, if the target class of the lock requester is subclass of the target class of the lock holder, there is no conflict. Otherwise, there are two cases as follows.

Subcase 3.1) There is no DC between the lock holder and lock requester

In this case, the possible conflict is detected on the target class of lock holder. For example, in Fig 3.c, let C5 and C6 denote the target class of a lock requester T_R and a lock holder T_H , respectively. In Fig. 3.c, the conflict is detected on the class C6.

Subcase 3.2) There is at least one DC between the lock holder and lock requester

In this case, the possible conflict is detected on the first DC of the lock holder through subclass chain of the lock holder. Let C1 and C6 denote the target class of lock requester T_R and a lock holder T_H , respectively. In Fig.3.d, the conflict is detected on the class C7, which is the first subclass of the lock holder.

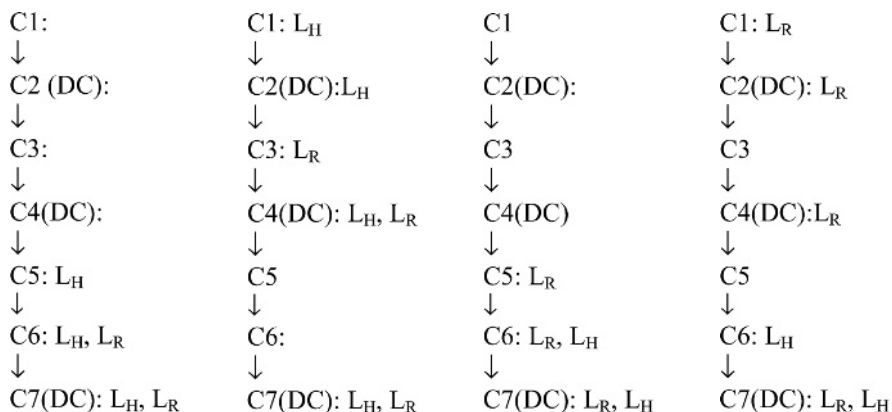


Fig. 3.a. A class hierarchy

Fig. 3.b. Access frequency for each class

Fig. 3.c. Results of DC assignment

Fig. 3.d. Results of DC assignment

Case 4) both are query transactions

Since a query transaction consists of read-only operations, there is no conflict between a lock requesting query transaction and a lock holding query transaction.

From cases 1), 2), 3) and 4), we can conclude that, for any lock requester, the possible conflicts with a lock holder are always detected. In turn, this means that our scheme satisfies the serializability since the proposed scheme is based on two-phase locking [1].

4 Performance Evaluation of the Proposed Scheme

In this Section, we will show that the proposed scheme performs better than the existing explicit locking. That is, assuming that access frequency by both OLTP transaction and query transaction is stable, we show that our scheme incurs less or equal number of locks than the explicit locking for any kind of access.

Claim) Assuming that access frequency by both OLTP transaction and query transaction is stable, the proposed scheme incurs less or equal number of locks than the explicit locking for any kinds of access.

Proof) We prove the above claim using induction on the number of classes checked in a class hierarchy. Let n denote the number of classes checked in a class hierarchy. Also, let $N(e)$ and $N(p)$ denote the number of locks required by the explicit locking and the proposed scheme, respectively.

1) $n=1$: $N(e) = N(p)$

2) $n=2$: assume that C_1 is a root and C_2 is a subclass of C_1 in a class hierarchy.

If C_1 is DC, $N(p) \leq N(e)$ otherwise C_1 would not be a DC by our DC assignment scheme. Similarly, if C_1 is non-DC, $N(e) = N(p)$.

Assume that the claim holds up to $n = K$

3) $n = K+1$

In this case, without loss of generality, we can assume that $(K+1)$ th class is a direct subclass of the class K .

Subcase 1) $(K+1)$ th class in non-DC

We prove that $N(p) \leq N(e)$. For locks required on the class $(K+1)$ th class for OLTP transaction, both schemes require the same number of locks. On the other hand, for locks required on the class $(K+1)$ th class for query transaction, the explicit locking requires locks on every subclass of $(K+1)$ th class as well as $(K+1)$ th class. However, our scheme requires locks on only DCs through subclass chain of the $(K+1)$ th class. For locks required for classes excluding $(K+1)$ th class, $N(p) \leq N(e)$ by induction assumption. Thus, $N(p) \leq N(e)$ for $n= K+1$.

Subcase 2) $(K+1)$ th class is DC.

Let $N(K+1: \text{non-DC})$ be the number of locks required when $(K+1)$ th class is non-DC for the proposed scheme. Then, $N(p) \leq N(K+1: \text{non-DC})$ otherwise, $(K+1)$ th class would not be DC. On the other hand, we prove that $N(K+1: \text{non-DC}) \leq N(e)$ as follows. For locks required on the class $(K+1)$ th class for OLTP transaction, both schemes require the same number of locks. On the other hand, for locks required on the class $(K+1)$ th class for query transaction, the explicit locking requires locks on every subclass of $(K+1)$ th class as well as $(K+1)$ th class. However, our scheme requires locks on only DCs through subclass chain of the $(K+1)$ th class. Since, for classes excluding $(K+1)$ th class, $N(p) \leq N(e)$ by induction assumption, $N(p) \leq N(e)$ for $n= K+1$.

From case 1), 2) and 3), we can conclude that our scheme incurs less or equal number of locks than the explicit locking for any kinds of access.

5 Conclusions and Further Work

In this paper, we present a locking-based concurrency control scheme for decision support in multimedia databases. Especially, our scheme aims at OODBs. In this work, transactions are classified into two groups, the typical OLTP transaction that reads and writes a few data items of database, and query transaction that is composed of read operation for decision support. Assuming that access frequency for a class hierarchy is

stable, our scheme performs better than the existing explicit locking for any kinds of access.

The immediate research issue is to find various characteristics of transactions in multimedia databases so that we adapt our scheme to such characteristics. In additions, since our scheme deals with only class hierarchy, we have a plan to develop a concurrency control scheme dealing with both class hierarchy and class composition hierarchy that is also a major concept of OODBs. We also extend our scheme to the Web databases that provide major source for decision support.

References

1. Bernstein, P., Hadzilacos, V. and Goodman, N.: Concurrency Control and Recovery in Database Systems, Addison-Wesley (1987)
2. Cart, M. and Ferrie, J.: Integrating Concurrency Control into an Object-Oriented Database System, 2nd Int. Conf. on Extending Data Base Technology, Venice, Italy, Mar. (1990) 363 - 377
3. Durham, M.: Data Mining, Prentice Hall, Upper Saddle River, NJ, USA (2003)
4. Garza, J. and Kim, W.: Transaction Management in an Object-Oriented Database System, ACM SIGMOD Int. Conf. on Management of Data, Chicago, Illinois, Jun. (1988) 37 - 45
5. Han, J. and Kamber, M.: Data Mining: Concepts and Techniques, Morgan-Kaufmann Press (2001)
6. Jun, W. and Gruenwald, L.: An Optimal Locking Scheme in Object-Oriented Database Systems, First Int. Conf. on Web-age Information Management 2000 (LNCS 1846), Shanghai, China, June (2000) 95-105
7. Khoshafian, S., Dasananda, S., and Minassian, M.: *The Jasmine Object Database: Multimedia Applications for the Web*, Morgan-Kaufmann Publishers, Inc., (1999).
8. Kim, H. and Park, S.: Two Version Concurrency Control Algorithm with Query Locking for Decision Support, ER 98 Workshop Advances in Database Technologies (LNCS 1552), Nov. (1998) 157-168.
9. Lee, L. and Liou, R.: A Multi-Granularity Locking Model for Concurrency Control in Object-Oriented Database Systems, IEEE Trans. on Knowledge and Data Engineering, Vol. 8, No. 1, Feb. (1996) 144-156
10. Malta, C. and Martinez, J.: Automating Fine Concurrency Control in Object-Oriented Databases, 9th IEEE Conf. on Data Engineering, Vienna, Austria, Apr. (1993) 253-260

Image Retrieval by Categorization Using LVQ Network with Wavelet Domain Perceptual Features

M.K. Bashar, Noboru Ohnishi, and Kiyoshi Agusa

Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
khayrul@agusa.i.is.nagoya.ac.jp
<http://www.ohnishi.i.is.nagoya-u.ac.jp/~khayrul/>

Abstract. Though most textile images are pattern dominant, there found a limited researches that focus on the pattern characteristics. In this study, we propose some perceptual features (directionality, regularity, symmetry) in the wavelet domain. Correlation among wavelet coefficients is the basis of the above features. In order to reduce searching time, we first categorize the database using supervised LVQ network. For each class, a class-vector is formed through averaging all the feature vectors in that class. The query key is first compared with class-vectors to come up with a category. It then performs similarity comparisons with the population of the selected category and retrieves relevant images. Users have also the provision to interact with the system if query fails to capture the relevant class. An experiment with a set of 300 curtain images shows the effectiveness of the proposed features compared to the well-known Gabor or discrete wavelet energy signatures.

Keywords: Retrieval, categorization, wavelet coefficient correlations, LVQ network.

1 Introduction

In the fashion, textile and clothing industry, handling a large amount of images is an important issue in various phases such as the designing, sourcing and merchandising phase. However, a limited number of investigations on the textile images were found in the literature. T. K. Lau et al. [7] proposed a system called “Montage”, where they use conventional second order statistics originally proposed by Haralick [15] as texture features. However, those features are computationally expensive and appear effective for microtextures only. Other established texture features [1], [2], [8], [5], [11], [12] like Gabor or wavelet energy statistics can not always extract perceptually meaningful attributes like regularity or directionality which reduces the so called semantic gaps for better retrieval. Perhaps Ballmelli et al. [4] attempted first on some wavelet domain features. However, they used computationally expensive region based features with tree-based classification system, which is sensitive to accurate feature threshold.

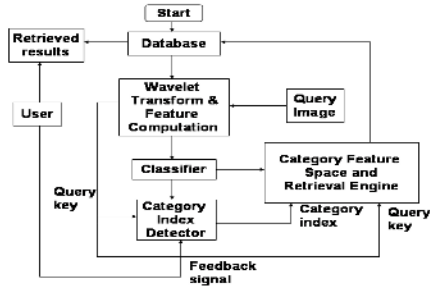


Fig. 1. A simple block diagram of the proposed system.

On the other hand, some recent studies [6], [9] emphasize on the edge-based measures for better retrieval.

We thus propose a simple system that uses three new wavelet domain features, namely directionality, regularity, and symmetry. Such features quantify textures in the way humans see them as regular, directional, symmetrical etc.

The remaining of this paper is organized as follows: section 2 gives the overview of our proposed system with details of the proposed directionality, regularity and symmetry features. Section 3 explains experimental details including discussion, while section 4 concludes the work with future directions.

2 Proposed Retrieval System

Our proposed system works on the signal representation approach of wavelet transform. The brief overview including the wavelet transformation technique is explained below.

2.1 Overview of the System

A simplified block diagram of our proposed retrieval system is shown in Fig. 1. First, the raw image database is inputted to the wavelet transform and feature computation (WTFC) block, which compute texture features and forward them to the classifier. Classifier categorizes entire database into several classes and stores respective feature vectors into the category-feature space. At the same time, class representative vectors are formed through averaging class members. A query image is applied to WTFC block, which produces query key. The query key and category vectors are applied to the category detector that detects the most similar category index. The query key and the category index are then applied to the category-feature space, where actual retrieval is performed. The retrieved images are verified by the user. If the user is not satisfied, a feedback signal is sent to the category detector. It then selects the second highest similar group and undergoes further retrieval until the scanning of all available groups.

2.2 Wavelet Domain Perceptual Features

Visual features, which approximate human visual perceptions, are very important. However defining such high-level features is not an easy job. In our study, we have formulated various features in the wavelet domain, mainly in the two perceptual directions, i.e., horizontal and vertical. In our implementation, we used Mallat's [10] fast 2D pyramid algorithm for discrete wavelet transform using Daubechies D4 wavelets through separate row and column filtering with downsampling. Thus, the obtained detail subbands can be generalized as $D_j^i(m, n)$ ($i = 0, 1, 2, 3; j = 0, 1, 2, \dots, \log_2^N - 1$), where $N \times N$ is the image size, i and j are the subband and scale indices, respectively.

Directionality (D). Directionality is usually defined on the edge direction histograms [14]. However, we defined it here on the directional wavelet subbands. L. Balmelli et al. [4] proposed a similar approach using region based correlation scheme. However, region consideration seems unnecessary in the wavelet domain. Instead, we can define it from the 1D correlation sequences, which can be expressed as a function row or column separation distances of the horizontal and vertical wavelet subbands as follows:

$$Cor(m, n) = \frac{\sum_{p=1}^N \{c_m(p) - \mu_{c_m}\} \{c_n(p) - \mu_{c_n}\}}{\sigma\{c_m(p)\} \sigma\{c_n(p)\}} \quad (1)$$

$$C(d) = \frac{1}{N_o(d)} \sum_{k=0}^{N_o(d)} Cor(kd, (k+1)d); N_o(d) = \lfloor \frac{N-1}{d} \rfloor \quad (2)$$

$$D_x(or D_y) = \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} C(d); d = 1, 2, 3, \dots, d_{max}. \quad (3)$$

Here m, n are row (or column) indices of wavelet sub-bands; $N, N_o(d)$ are the number of rows (or columns), and row-pairs (or column-pairs) of sub-bands at specified distance, d ; μ_c is the mean and $\sigma(\cdot)$ is the standard deviation of the coefficients (c_m 's) corresponding to the m -th row or column. Theoretically, we can use $d_{max} = N - 1$. However, we used 16 as a d_{max} value in our current implementation.

Regularity (R). Regularity can be computed from the correlation sequences of detail subbands. Correlation sequence for a region of size s (number of rows or columns) is defined by

$$c_s(m) = \frac{1}{s} \sum_{n=m+1}^{m+s} Cor(m, n) \quad (4)$$

where $Cor(m, n)$ is the normalized correlation of coefficients between two rows (or columns) of subbands as in Eq. 1 and $s = 2^l$, where $l = 1, 2, \dots, 5$. The

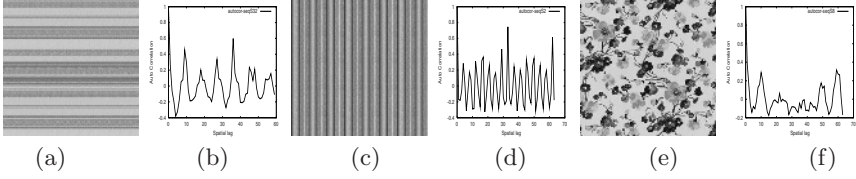


Fig. 2. Periodicity analysis; (a,c,e) Sample curtain images (hpatt15, vpatt1, fpatt5); (b,d,f) Autocorrelation curves from horizontal subbands(level-1).

autocorrelation of the above sequence may be obtained by:

$$AC_s(k) = \frac{\sum_{m=0}^{N-1} (c_s(m) - \mu_{c_s}) \times (c_s(m+k) - \mu_{c_s})}{\sum_{m=0}^{N-1} (c_s(m) - \mu_{c_s})^2} \quad (5)$$

Here μ_{c_s} is the mean of $c_s(m)$ and $k = 1, 2, \dots, N/2$. The above sequence, $AC_s(k)$ may provide approximate periodicity at a particular lag, where it shows the local peak value. We may expect a peak value in $AC_s(k)$ for each region size s . So the peak autocorrelation as a function of region size s can be expressed as

$$PA(s) = \max_k AC_s(k) \quad (6)$$

The following parameters can then be defined from the above $AC_s(k)$ and $PA(s)$ as measures for regularity.

$$r_s = \arg \max_s PA(s) \quad (7)$$

$$r_m = \max_s PA(s) \quad (8)$$

$$r_p = K_1 \arg \max_k AC_{r_s}(k) \quad (9)$$

Here r_s and r_p are the approximate primitive size and periodicity, where maximum value of autocorrelation, r_m , occurs. K_1 is a constant, whose value depends on wavelet scale. Fig 2 shows $AC_s(k)$ plots for the mentioned images. However, autocorrelation may produce some spurious peaks even though $c_s(i)$ does not show any periodicity (e.g., fpatt5 in Fig. 2). In order to minimize such unexpected peaks, we have performed a thresholding operation on $c_s(i)$ before computing autocorrelation.

Symmetry(S). As symmetry measure, we have considered only mirror symmetry, namely horizontal and vertical symmetry. A soft symmetry measure can be computed from the binary edge map $G(m, n)$, obtained by thresholding an edge image, $E(m, n)$. For a decomposition level (say $j = 1$), $E(m, n)$ can be derived from the horizontal and vertical subbands (of size $N \times N$) by

$$E(m, n) = \sqrt{(D_1^1(m, n))^2 + (D_1^2(m, n))^2} \quad (10)$$

Similar edge images are computed for other resolution levels. Thus, an approximate normalized horizontal symmetry can be defined as the total number of symmetric edge point-pairs around the n -th pixel in a row of $G(m, n)$:

$$S_{nor}^H(m, n) = \frac{2}{N} \sum_{a=1}^{N/2} edge(m, n, a) \quad (11)$$

where

$$edge(m, n, a) = \begin{cases} 1 & \text{if } G_j(m, n+a) = G_j(m, n-a) = 1; \\ & \text{where } a = 1, 2, \dots, N/2. \\ 0 & \text{Otherwise} \end{cases}$$

For the m -th row, the most probable symmetry score can be obtained by taking the maximum value of the above normalized symmetry by

$$SS(m) = \max_n \{S_{nor}^H(m, n)\}; m = 1, 2, \dots, N. \quad (12)$$

$$SL(m) = \arg \max_n \{S_{nor}^H(m, n)\}; m = 1, 2, \dots, N. \quad (13)$$

Now using the Eqs. 12 and 13, the following parameters (symmetry score, symmetry location and symmetry value) can be defined as measures for image symmetry:

$$SS_H = \frac{1}{N} \sum_{m=1}^N SS(m), \quad SL_H = \frac{1}{N} \sum_{m=1}^N SL(m), \quad SV_H = K \frac{\sum_m SS(m)}{\sigma(SL(m)) + \delta}$$

Here K is constant and δ is a small real number, which avoids infinite condition for the symmetry value. Similar parameters can be obtained for vertical symmetry at all resolution levels.

To compare the performance of the proposed features, we used the most widely used Gabor energy (GE) signatures [16] defined by

$$GE_{ij} = \sum_{m=0}^M \sum_{n=0}^N |g_j^i(m, n)| \quad (14)$$

where $g_j^i(m, n)$ is the i -th even Gabor channel of j -th decomposition level. M, N are the channel dimensions. Similar formulation can be used for wavelet energy (WE) feature.

3 Experimental Results

3.1 Data-Set

In our experiment, we have used 300 curtain images from SANGETSU (Japan) company (<http://www.sangetsu.co.jp>). The set contains two different image sizes, i.e., 256×256 and 240×240 , respectively. These color images are first transformed using CIE $L^*a^*b^*$ transformation. Only the luminance channel (L) is used in our study. We have, manually, examined a maximum of six (6) pattern categories in the data-set. The categories may be designated as (i) Horizontal

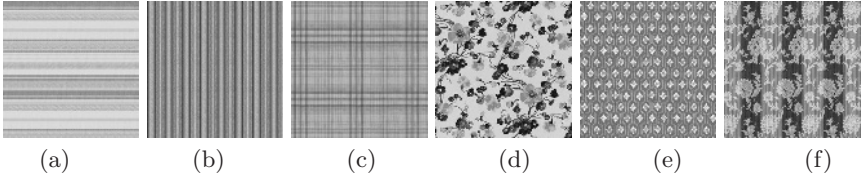


Fig. 3. Some sample curtain images of six perceptual categories; **Horizontal:** (a) hpatt16, **Vertical:** (b)vpatt1, **Cross:** (c)xpatt10, **Non-regular:**(d) fpatt5, **Regular:**(e)rpatt1, and **Mixed:**(f) mpatt5.

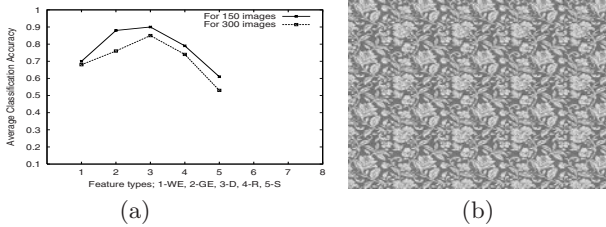


Fig. 4. Accuracy and Query: Average classification accuracy (a) for 150 and 300 images, (b) Query image for retrieval(fpatt45).

Table 1. Average classification accuracy for the features

Database size	Average classification accuracy (%)				
	WE	GE	D	R	S
150 images	70.00	88.00	90.00	78.66	60.66
300 images	68.00	76.00	84.66	74.00	53.33

(HH), (ii) Vertical(VV), (iii) Cross (XX), (iv) Non-regular (NR), (v) Regular (RR), and (vi) Mixed (MM). Fig. 3 shows some sample curtain images.

The well-known LVQ network [13] is used to categorize the database. We chose 80 random samples (13 per category) out of 300 images for training the LVQ network. Fig. 4 shows the average classification accuracy for all features. Clearly, the proposed directionality attains the best performance. A small variation in accuracy is observed with increasing size of the database as shown in table 1 and Fig. 4.

3.2 Retrieval

Our system performs retrieval from the categorized images as explained in section 2. Figs. 5 and 6 show the first 12 retrieved images in response to fpatt45 query image as shown in Fig. 4(b). These figures show the relevance order of

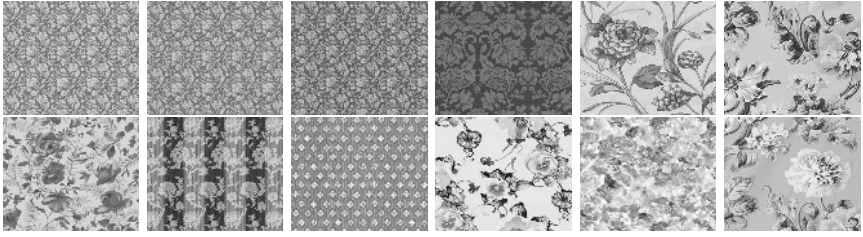


Fig. 5. Retrieval results for GE feature. First 12 retrieved images(row-wise) are xpatt45, fpatt44, fpatt46, fpatt1, fpatt47, fpatt50, fpatt13, mpatt5, rpatt2, fpatt8, fpatt22, fpatt49 (No feedback, irrelevant-2 (mpatt5, rpatt2)). Note that GE feature can obtain 42 out of 65 images.

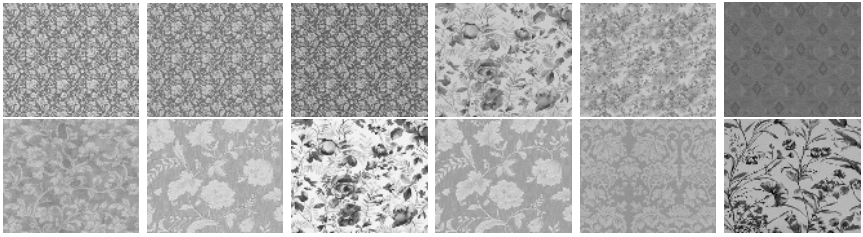


Fig. 6. Retrieval results for D feature. First 12 retrieved images(row-wise) are fpatt45, fpatt44, fpatt46, fpatt12, fpatt30, rpatt27, fpatt39, fpatt56, fpatt10, fpatt57, fpatt2, fpatt7 (No feedback, irrelevant-1 (rpatt27)). Note that D feature can obtain 65 out of 65 images.

the retrieved images for the directionality and GE features. Results for other features are not shown for space limitation.

Two well-known metrics, namely *precision* and *recall* [3], are used to evaluate our system. A precision-recall curve is usually based on 11 standard recall levels (0 %, 10%, 20%, ..., 100%). Since retrieval depends on classifier and feature performance, we cannot obtain 100 % recall rates for all queries. Hence, the averaging is done on the common maximum recall rate for the query-set for each feature. Retrieval effectiveness is always evaluated by averaging interpolated precisions for more than one query. Fig. 7(a) and (b) show the recall, precision graphs corresponding to fpatt45 query, while Fig. 7(c) and (d) indicate the interpolated P-R (IPR) graphs for two sets of queries (one sample per set per category) from databases of 150 and 300 images. The above graphs show that directionality feature is quite promising. It attains the highest precision for 150 images. For 300 images, though GE and WE features take a lead over D, it again exceeds them at about 40 % recall rate. Moreover, it always achieves the maximum recall rate indicating its robustness in retrieval.

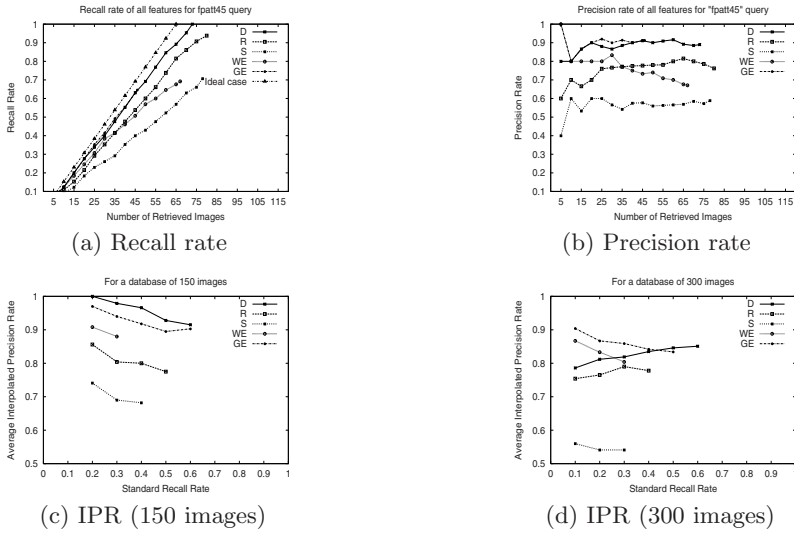


Fig. 7. The performance comparison among various features (D, R, S, WE, GE). (a) Recall, (b) Precision for query:fpatt45; Interpolated PR curve for 6 queries from a database of (c) 150 images and (d) 300 images.

3.3 Discussion

The convergence criterion of the LVQ network is an important factor to consider. At present, as a rule of thumb, the total iteration is fixed experimentally to ten times the total category in the database. Current IPR's are based on a single query-set that consists of one sample per set per category. However, usage of more queries per set per category will enhance the reliability of IPR curve. Moreover, a type of pattern complexity measure has to be designed in future for better pattern characterization. We are also planning to develop an automatic categorization technique for better applicability.

4 Conclusion

A novel approach for retrieving textile-curtain database is proposed based on wavelet-domain perceptual features, namely regularity, directionality and symmetry. Among the proposed features, directionality achieves the highest performance in terms of precision, recall and speed. However, a pattern complexity measure has to be formulated to capture more complex patterns. We will also develop an automatic categorization technique for a future intelligent system.

Acknowledgements. We would like to thank Dr. H. Kudo, T. Matsumoto, Y. Takeuchi, and T. Yamamura for their invaluable suggestions during the work in progress.

References

1. Bashar M. K., Matsumoto T., Ohnishi N.: Wavelet Transform-based Locally Orderless Images for Texture Segmentation. *Pattern Recognition Letters*. **24(15)** (2003) 2633–2650
2. Bashar M. K. and Ohnishi N.: Integrating Cortex Transform and Brightness Based Features for Multi-texture Classification. *The J. Inst. Image Info. and Television Engrs.* **56(11)** (2002) 1769–1778
3. Jones K. S.: *Information Retrieval Experiment*. Butterworth and Co. (1981)
4. Balmelli L. and Mojsilovic A.: Wavelet Domain Features for Texture Description, Classification and Replica Analysis. *Proc IEEE Intl. conf. on Image Process.* **4** (1999) 440–444
5. Tian Q., Sebe N., Lew M.S., Loupias E., and Huang T.S.: Image Retrieval Using Wavelet-based Salient Points. *J. Electronic Imaging*. **10(4)** (2001) 835–849
6. Cheng-Hao Y. and Shu-Yuan C.: Retrieval of translated, rotated and scaled color textures. *Pattern Recognition*. **36** (2003) 913–929
7. Lau T. K. and King I.: Montage: An Image database for the Fashion, Textile, and Clothing Industry in Hongkong. *Proc. of the 3rd Asian Conference on Computer Vision*. **1** January 4-7 (1998) 410–417
8. Manjunath B. S. and Ma W. Y.: Texture features for browsing and retrieval of large image data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. **18(8)** (1996) 837–842
9. Kubo M., Aghbari Z., OH K. S., and Makinouchi A.: Image Retrieval by Image Features Using Higher Order Autocorrelation in a SOM Environment. *IEICE Trans. on Information and System*. **E86-D(8)** (2003) 1406–1415
10. Mallat S.: The theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.* **11(7)** (1989) 654–693
11. Suematsu N., Ishida Y., Hayashi A., and Kanbara T.: Region-based Image Retrieval using Wavelet Transform. *Proc. of the 15th Int. Conf. on Vision Interface*. May 27-29 Calgary, Canada(2002) 9–16
12. Paquet A. H., Zahir S., and Ward R. K.: Wavelet Packets-Based Image Retrieval. *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*. May13-17, Florida, USA (2002) Paper no.1242
13. Kangas J. A., Kohonen T., and Laaksonen J. T.: Variants of self organizing maps. *IEEE Trans. on Neural Networks*. **1(1)** (1990) 93–99
14. Tamura H., Mori S., and Yamawaki T.: Texture features corresponding to visual perception. *IEEE Trans. on Systems, Man, and Cybernetics*. **8(6)** (1978) 460–473
15. Haralick R. M., Shanmugam K., and Distein I.: Texture features for image classification. *IEEE Trans. Syst. Man. Cybern.* **3(6)** (1973) 610–621
16. Nestares O., Navarro R., Portilla J., and Taberbero A.: Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. *J. Electronic Imaging*. **7** (1998) 166-173

A Content-Based News Video Browsing and Retrieval System: NewsBR

Huayong Liu¹ and Zhang Hui²

¹ Department of Computer Science, Central China Normal University,
Wuhan 430079, Hubei, PR China
hyliuwuhee@hotmail.com

² Finance Department of Business School, Wuhan University,
Wuhan 430072, Hubei, PR China
zhanghui994@sohu.com

Abstract. An advanced content-based news video browsing and retrieval system, NewsBR, is proposed in this work. The system is built on high-accuracy news story segmentation and topic caption text extraction. Its main features include category-based news story browsing, key-frame-based video abstract and keyword-based news story retrieval. In the paper, news story segmentation and topic caption text extraction, as well as content-based video browsing and retrieval, are addressed in detail. The system is helpful and effective for the overall understanding of the news video content.

1 Introduction

As more and more video libraries are available to homes and offices through the Internet, it becomes increasingly important to design systems for efficiently browsing and retrieving the video content. What does content mean and how to characterize video content are two big problems for video information researchers. It is generally accepted that content is too subjective to be characterized completely because it is often concerned about objects, background, domain, context, etc. This is one of the main reasons why the problem of content-based access is still largely unsolved. Ref.[1] has presented a video browsing method, which is conducted through a VCR-like interface. It is tedious and time-consuming, and is not concerned about high-level semantic content understanding. Ref.[2] has presented a video's structured browsing and querying system called videowser, but it also has not realized effective content-based retrieval and just considers the image analysis, so it is not a really content-based video browsing and retrieval system.

Though much research work has been made towards developing automatic video searching system in recent years, however, because of the numerous video program variations, it is still a very difficult work to design a general-purpose system for all types of video programs. In this paper, we focus on TV news programs as a particularly important category of video programs and design a content-based news video browsing and retrieval system, NewsBR, which is convenient for

users to fast browsing and retrieving news video. Combining audio-visual features and caption text information, the system automatically segments a complete news program into separate news stories. Then using automatically extracted text caption and results of speech recognition as index files, NewsBR supports keyword-based news story retrieval. The system also supports category-based news story browsing and generates key-frame-based video abstract for each story.

2 Story Segmentation of News Video

A news program usually consists of dozens of news stories. To browse and retrieve these stories we first need to segment the continuous news programs into individual news stories. The NewsBR system carries out the news video story segmentation through three functional modules, the shot boundary detection module, the caption text detection module and the silence clip detection module, as shown in Fig. 1.

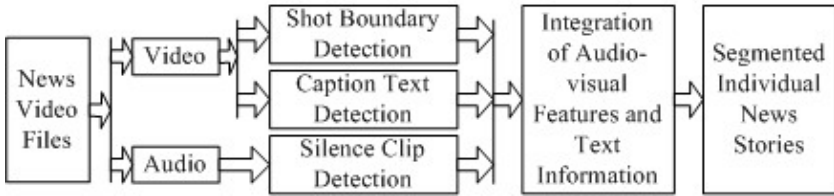


Fig. 1. The framework of the news video story segmentation scheme in the NewsBR system

2.1 Shot Boundary Detection

The shot boundary detection module function is to identify the scene changes or the boundaries between camera shots in a video stream. Because of spot and time limitations of news video in shoot and edition, the abrupt shots of news video obtain a rate of 90news programs. Even if they occur in the main body of news programs, they generally occur inside of the news story and do not locate at news story boundaries. So in shot boundary detection module we just only consider the detection of abrupt transition because there should be abrupt transition between two consecutive news stories to a great extent.

In order to detect shot boundaries, the method of X^2 histogram matching mentioned in Ref.[3] is used in this paper to measure the content change between each contiguous frame pairs. The expression is written as follows,

$$d(I_{t+1}, I_t) = \sum_{k=1}^n \frac{(H_{t+1}(k) - H_t(k))^2}{H_t(k)}. \quad (1)$$

where $d(I_{t+1}, I_t)$ represents the color histogram difference of frame image I_{t+1} and I_t , and H_{t+1} , H_t is the respective histogram, n is the size of histogram array. Using a data set of 1026 news video shots, we obtain an accuracy rate of 92% and a recall rate of 97%.

2.2 Topic Caption Text Detection

The next step is to detect the topic caption texts of each news story. For every news story of TV news programs, a row or several rows of captions must appear to express the meaning of this story in the start or middle. They are added by late execution of new programs, generally appear in the specific position (for example, bottom-left part of TV screen) and last several seconds. These captions are defined as topic captions, and the frame containing the topic captions is defined as topic-caption frame. The appearance and the disappearance of these captions are defined as a text event.

In order to detect the topic-caption frames, we present a new method called the specific region text detection algorithm based on threshold. Two frame difference sequences are calculated: the whole frame difference sequence and the part frame (the 1/4 field of screen bottom where text events appear) difference sequence. If the part frame difference is higher than the global threshold of text event and the whole frame difference is lower than the global threshold of abrupt shot transition, there is a text event. If the part frame difference is higher than the global threshold of text event and the whole frame difference is also higher than the global threshold of abrupt shot transition, there is an abrupt transition.

Some else captions will added in the same region except the topic captions in late execution of news programs. For example, the dialogue captions between a reporter and an interviewee. In order to avoid wrong detection, we should distinguish the topic-caption frames from these frames containing else captions. The two horizontal edges of caption text region are detected using edge detection algorithm to confirm the topic-caption frame. Most present edge detection algorithms are mature and effective, so the algorithm described in Ref.[4] is used in this paper.

Based on analysis above, the topic-caption clips $TC(n)$ can be obtained, and it can be expressed as $TC(n) = [V_{s_n}, V_{e_n}]$, $n = 1, 2, \dots$, and $V_{s_n} < V_{e_n}$. V_{s_n} and V_{e_n} represent the corresponding frame number of the topic captions start and end respectively. We choose the frame at the position $\text{INT}((V_{s_n} + V_{e_n})/2)$ as the topic-caption frame of the n th story and $\text{INT}(a)$ is defined to get the integrate part of a . Then the array $Ftc(n)$ of topic-caption frames sequence can be obtained, which is expressed as $Ftc(n) = \{Ftc_1, Ftc_2, \dots, Ftc_n\}$.

2.3 Silence Clip Detection

According to lots of experimental observation of news video, we find out that a relatively long silence clip must exist in two continuous news story boundary, so it is of great significance to find these silence clips. The frame containing silence clip is defined as silence-clip frame. In our system, the audio stream is defined

as two layers that include audio frame and audio clip according to the temporal relationship. We choose an audio frame as about 20ms, and continuous 30 audio frames as an audio clip.

For audio signals, short-time energy is an essential parameter for distinguishing silence clips from non-silence clips. It is evident that the short-time energy values of silence clips are remarkably lower than those of non-silence clips. The short-time average zero-crossing rate (ZCR) is another effective measurement to differentiate silence clips from non-silence clips, as the silence clips have much smaller ZCR values than the non-silence clips.

Both energy and ZCR measures are used to detect silence clips. If the short-time energy function value of an audio frame is lower than the defined threshold T_e and its short-time zero-crossing rate is lower than another defined threshold T_{zcr} , then the audio frame is indexed as a silence frame SF , also define $SF^S = 1$. For audio clip $AC_{V_{t1}, V_{t2}}^i$, V_{t1} and V_{t2} represent the start and the end frame number of AC^i respectively, $AC_{V_{t1}, V_{t2}}^{iS} = \sum_{V_t=V_{t1}}^{V_{t2}} SF_{V_t}^S$, if $AC_{V_{t1}, V_{t2}}^{iS} \geq \beta$, then $AC_{V_{t1}, V_{t2}}^i$ is defined as the i th silence clip, and $Esc_{V_{t1}, V_{t2}}^i$ is defined as the silence event. The following parameters are chosen, $T_e = 10000$, $T_{zcr} = 0.02$, $\beta = 24$. Based on analysis above, we can get a series of silence clips $SC(n) = \{SC_{V_{s1}, V_{e1}}^1, SC_{V_{s2}, V_{e2}}^2, \dots, SC_{V_{sn}, V_{en}}^n\}$, $n = 1, 2, \dots$, and $V_{sn} < V_{en}$. V_{sn} and V_{en} represent the corresponding start and end frame number of SC^n respectively.

2.4 Segmentation of News Stories

We can get the important audio-visual cues and text information from above to segment news stories. For a news story NS_k , it has only one topic-caption frame Ftc_k , and there should be one or more silence clips before and after appearance of Ftc_k in temporal axis. These silence-clip frames at the story boundaries should be abrupt transition except that a whole news story is inside of an anchorperson shot and it has not live broadcasts, so we can detect shot transition of these specific silence-clip frames between two consecutive topic-caption frames to locate the story boundaries. We define E_{at} as the abrupt shot transition event.

Based on the above analysis, we present the approach of locating news stories boundaries as follows,

1. For $SC_{V_{sn}, V_{en}}^n$, if $[V_{sn}, V_{en}] \in [Ftc_k, Ftc_{k+1}]$, $Esc_{V_{sn}, V_{en}}^n \cap E_{at} \neq \Phi$, then the frame at the position $\text{INT}((V_{sn} + V_{en})/2)$ is chosen as the story boundary between NS_k and NS_{k+1} .
2. For $SC_{V_{si}, V_{ei}}^i$ and $SC_{V_{sj}, V_{ej}}^j$, if $[V_{si}, V_{ei}] \in [Ftc_{k-1}, Ftc_k]$, $[V_{sj}, V_{ej}] \in [Ftc_k, Ftc_{k+1}]$, $Esc_{V_{si}, V_{ei}}^i \cap E_{at} \neq \Phi$, $Esc_{V_{sj}, V_{ej}}^j \cap E_{at} \neq \Phi$, it shows that the news story NS_k is inside of one anchorperson shot and there is no abrupt transition around news story NS_k . We calculate $\lambda = V_{s\lambda} - V_{e\lambda} = \max\{V_{si} - V_{ei}\}$, $[V_{s\lambda}, V_{e\lambda}] \in [Ftc_{k-1}, Ftc_k]$, and choose the frame at the position $\text{INT}((V_{s\lambda} + V_{e\lambda})/2)$ as the story boundary between NS_{k-1} and NS_k .

Then calculate $\Theta = V_{s_\Theta} - V_{e_\Theta} = \max\{V_{s_j} - V_{e_j}\}$, $[V_{s_\Theta}, V_{e_\Theta}] \in [Ftc_k, Ftc_{k+1}]$, and choose the frame at the position $\text{INT}((V_{s_\Theta} + V_{e_\Theta})/2)$ as the story boundary between NS_k and NS_{k+1} .

Using test data set, CCTV news selected from our video database randomly, which lasts one and a half hours or so in total and contains 135,400 frames, we obtain an accuracy rate of 85.8% and a recall rate of 97.5% for news video story segmentation.

3 Topic Caption Text Extraction

The topic caption texts in each story frame are valuable descriptions of the story content. Automatic extraction and recognition of these caption texts provide an efficient approach to annotate story contents.

A text segmentation algorithm is designed based on edge detection for extracting and recognizing Chinese captions in news video programs. The objects are the still images in BMP format that contain the topic caption texts, which are obtained from the results of caption text detection module. We use two or-

inary Sobel arithmetic operators. $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ is used to detect the horizontal

edge, and $\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ is used to detect the vertical edge, so we can get the rect-

angle region that contains the texts. Because the texts in video flows mostly are arrayed horizontally, so if we project the image onto horizontal direction (Y axis), very steep peak value appears in the corresponding area certainly, and every peak's width is equal to every text's height. In the same way, if we project the pixels of every row of texts in the corresponding Y-axis area onto vertical direction (X axis), we can get its width. Then we call every region marked in this way as a block, and we call this approach as projection method. The experiments prove this method is effective and fast. According to the block extracted in above steps, we can extract the corresponding region from the original image. Because the gray scale difference between the texts and its background is great in the part region that contains the texts, so we can process the region by binarization based on threshold. Then the text region is transmitted to the recognition core of commercial OCR software (we use Thocr 7.5 SDK demo version of Tsinghua Ziguang), we get the final recognition results. An example is shown in Fig. 2.

4 Content-Based News Video Browsing and Retrieval

According to analysis of news video above, we present content-based news video browsing and retrieval. Now there are lots of researches about this field, Ref.[5] has proposed the method of content-based retrieval and browsing, including key-frame-based retrieval, shot-based retrieval and key-frame-based hierarchical

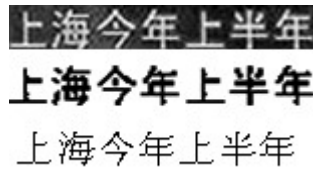


Fig. 2. An example of caption text recognition results

video browsing. Ref.[6] has presented the method of automatic content-based retrieval of broadcast news using information retrieval. As for this paper, we combine the method of content-based browsing news stories based on category, video abstract based on key frames and the method of querying by inputting keywords.

Browsing video hierarchically is an important method to obtain video content. Some keywords are predefined in a small news vocabulary database that are built according to our observation results for a month of CCTV news programs and the database can be expanded in future work. We classify news stories into eight categories including politics, economy, science, culture, sports, entertainment, etc. And every category has a group of keywords, which have different weight values. In the one hand, we integrate the speech recognition engine that is programmed with the API functions in the NewsBR system. The API functions are provided by the speech recognition development kit, Microsoft Speech SDK 5.0 of Microsoft corps. This engine recognizes the anchorperson voices in every news story start. In the other hand, the topic caption texts recognition results are obtained according to Section 2. Based on keywords formed by the two kinds of text recognition results of every story, which are predefined in news vocabulary database, we calculate the sums of these keywords' weight values in every predefined category. The category whose sum is highest is considered as the category of this story, so as to provide more convenient measure to browse news stories. We can browse every single news story in different categories through the graphical user interface.

Key frame can let users know general meanings of a news clip quickly, and now there are lots of algorithms that introduce how to extract key frame such as Ref.[7][8]. Considering the specification of TV news programs, we select two kinds of frames as key frames in the key frame extraction module of our system. The first is the frame that contains the topic caption texts, and users can master general meanings of the news story by watching the texts. The second is the frame of whole shots in MPEG flows of every news story, which is closest to the middle point in temporal space. The key frames form the video abstracts and they are stored in JPEG format in order to save system spending. So we can also browse the video abstracts directly to learn general meanings of news stories.

Querying by users' inputting keywords is an efficient method for video retrieval. The two kinds of recognized texts described above form full-text search indices, when users input keywords whatever they thought about such as "Lipeng" or "Investment" to query news video, and then the system can

provide the exact video abstracts and play the corresponding news stories in different categories.

5 The Interface of NewsBR and Its Functions

Fig. 3(a) shows the interface of our system. It consists of five sections, the left display window, the right display window, the story results window, the key-frame-based video abstract window and the key word inputting query window. User is able to input news video files, and set the start and end of the video files. Clicking the play button in the new video control dialog, the story segmentation of the inputted video files is automatically finished. The results of segmented stories and extracted key frames are shown in the story results window and the key-frame-based video abstract window respectively. The left display window is used to play a video file in realizing story segmentation. And the right display window is used to play a story selected from the segmented results or to show a key frame selected from the extracted key frames. User is able to input keywords to search news stories, for example, the word “Lipeng”, and then the retrieval result is shown in the retrieval results dialog, as shown in Fig. 3(b). Click the first key frame and the corresponding news story can be played in another dialog box.

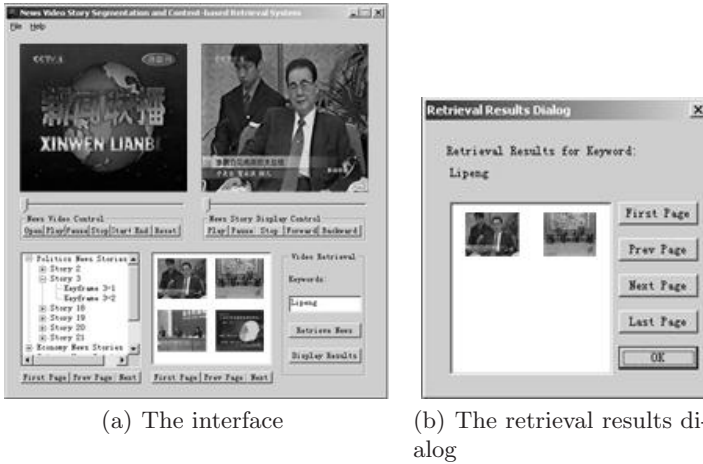


Fig. 3. The interface of our system and the retrieval results dialog

6 Conclusions

Content-based video browsing and retrieval for video flows is a hot spot in the recent researches of video database. This paper develops a system, called NewsBR,

to realize fast and efficient news video browsing and querying based on accuracy story segmentation and topic caption texts extraction. The system is designed for parsing TV news, but its integration strategy of audio-visual cues, the analysis of text event detection, as well as the methods of content-based video browsing and retrieval can also be applied to the scene segmentation and video retrieval of other video types in future work.

References

1. Y.H.Chang, D.Coggins, D.Pitt, D.Skellem, M.Thapar, and C.Venkatraman: An Open-Systems Approach to Video on Demand. *IEEE Communications Magazine* (1994) 68–80.
2. Wu Lingqi, Li Guohui: Video's Structured Browsing and Querying System: Videowser. *Mini-Micro System* (2001) 112–115.
3. A. Nagasaka and Y. Tanaka: Automatic Video Indexing and Full-Video Search for Object Appearance. In: *Proceedings of Second Working Conference on Visual Database Systems* (1991) 113
4. Ramin Zabih, Justin Miller and Kevin Mai: Feature-Based Algorithms for Detecting and Classifying Scene Breaks. In: *Proceedings of Fourth ACM Conf. on Multimedia* (1995) 189–200
5. Zhang H J, Low C Y, Smoliar Stephen W and Wu J H: Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution. In: *ACM Multimedia 95-Electronic Proceedings* (1995) 15–24
6. Martin G. Brown, Jonathan T.Foote, Gareth J.F.Jones, Karen Sparck Jones and Steve J.Young: Automatic Content-Based Retrieval of Broadcast News. In: *ACM Multimedia 95-Electronic Proceedings* (1995) 35–43
7. W.Wolf. Key Frame Selection by Motion Analysis: In: *Proceedings of IEEE International Conference Acoustic, Speech, and Signal Proceeding* (1996) 1228–1231
8. P.O.Gresle and T.S.Huang: Gisting of Video Documents: A Key Frames Selection Algorithm Using Relative Activity Measures. In: *Proceedings of the Second International Conference on Visual Information Systems* (1997)

A News Video Browser Using Identical Video Segment Detection

Fuminori Yamagishi¹, Shin'ichi Satoh², and Masao Sakauchi²

¹ The University of Tokyo, Japan,

² National Institute of Informatics,

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

{fuminori, satoh, sakauchi}@nii.ac.jp

Abstract. Structuring video archives is an important task to make full use of them. In broadcast news videos, some kind of video segments appears repeatedly. Those segments, which we call identical video segments, are useful for structuring or analyzing video database on our observation. In order to confirm that, we searched identical video segments from 350 hours-long broadcast news video archive and implemented a news video browser using results of identical video segment detection. We found that the distribution of those segments can be used to extract important topics from the news archive, and also found that time intervals of appearance of the same video segment can be used for classify identical video segments.

1 Introduction

With current technologies, people can have many large broadcast video archives from TV programs. This enables extracting more information or using archives more usefully than ever, by directly using high-dimensional information like video or audio data. To make full use of video archives or to enable content-based video access to the archives, people need efficient methods to structuring or analyzing them.

In video archives, there is some video segment that appears repeatedly, and those video segments can be used for analyzing the video archives [1]. As a basic component of content-based video access, researchers tend to use feature-based similarity search for images and videos. However, several papers have recently been published which state that similarity search is getting noisier and useless as the image/video archives are getting larger, instead, searching the “identical” image/video is becoming useful [2,3].

In this paper, we are especially focusing on structuring news video archives. There are many news channels that broadcast news programs 24 hours a day, and there are also broadcast stations which broadcast several hours of news program a day. Those broadcast news videos occasionally have repeatedly used video segments. Examples of identical video segments detected from an actual news video archive are shown in Fig. 1. Images like Fig.1 (a) or (b) appears almost everyday in a collection of a news program. Other images in Fig. 1 are appeared repeatedly

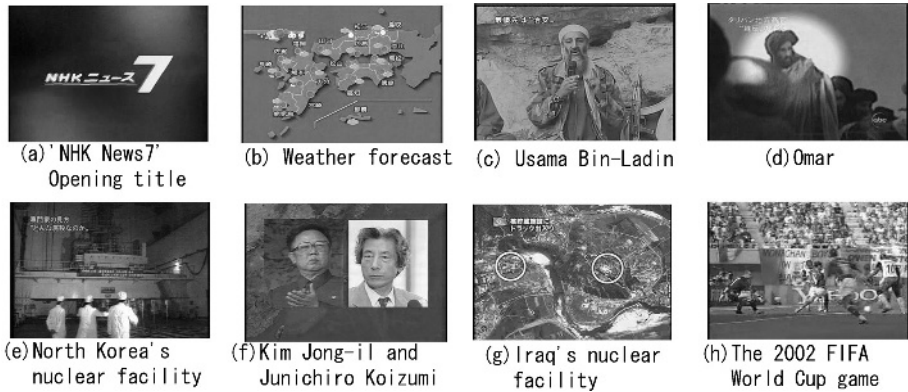


Fig. 1. Examples of identical video segments from a news archive

because they are used in well-noticed topics, which broadcasted repeatedly for several days. One of the reasons of this repetition is some kind of reference videos are usually inserted in news videos, and creating or newly acquiring reference video usually costs high.

We noticed that most parts of video archives that share the same video segment has some relations, and such segments tends to be classified depending on their distribution in the broadcast time scale. In this paper we search identical video segments in a news video archive and we implement a news video browser that uses search results of identical video segments, and investigate how the method of identical video segment detection can be used for structuring or extracting valuable information from news video archive.

2 Identical Video Segment Detection

Identical Video Segment Detection is a method that searches repeatedly appeared video segments in video archive. When you search in several ten hours of news video archive, you will find that several video segments appear repeatedly in the video archive. We claim that these video segments can convey useful and important topics and contribute structuring the video archive. Several example pairs of frame image that should be treated as identical are shown in Fig. 2.

Our method of identical video segment detection is carried out by the following algorithm.

1. Decompose video stream into shots.
2. For every pair of shots, judge whether they share at least one “identical frame image pair”. “Identical frame image pair” is a pair of frame image of which normalized cross correlation is larger than some threshold (eg. 0.90).
3. If the shot pair shares at least one “identical frame image pair”, they are identical video segment.



Fig. 2. Example pairs of frame image that we treat as identical

Every insertion of identical video segment is not always occurred at shot boundaries, frame images should be compared frame by frame.

When an editor of news programs uses the same video segment again and again, it is likely that he changes telops or some kind of post productions added on the repeated video segment for each use. But we want to search those video segments regardless of those post edits, since those edits don't invalidate the relations of the parts that share the same video segment. In order to manage that, we decided to set the correlation threshold to slight less than 1. We set it 0.90 actually according to our observation.

Although the algorithm showed above is simple enough, but since it involves frame-by-frame comparison, it costs enormously high when you search in large video archive. For example, if we search identical video segments in a 100 hours-long video archive, the processing time with a typical PC system is about 50000 years, which obviously is intractable. In order to extract meaningful results, we have to apply to large video archive, so some acceleration method has to be used.

As a acceleration method, we use pre-filtering using normalized intensity histogram. The outline of the method is following.

1. Extract a low dimensional histogram (16 dimensional normalized intensity histogram).
2. Filter out frame image pair of which histogram distance is larger than the threshold that is pre-decided.
3. Compute normalized cross correlation for the remaining pairs.

The normalized intensity histogram is confirmed by our experiment [4] that it can filter out not "identical frame image pair" while there are almost no false drops. Computing distances of 16-dimensional histogram is much less than computing

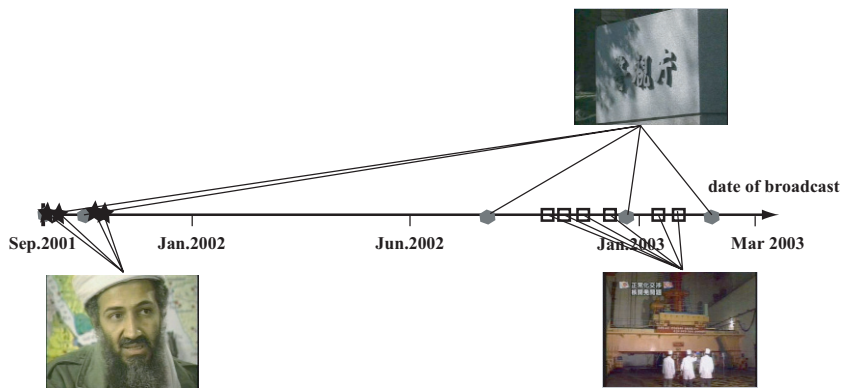


Fig. 3. Distribution of dates when each identical video segment broadcasted

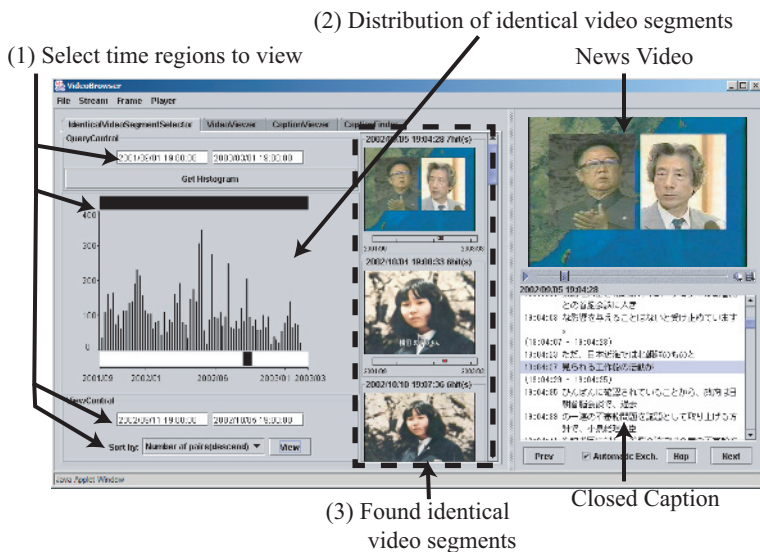


Fig. 4. Identical Video Segment Browser

normalized cross correlations, so searching identical video segment can be carried out efficiently.

In order to investigate the effect of Identical Video Segment Detection, we applied the method of Identical Video Segment Detection to an actual news archive. The archive we used is consisted of about 700 days video footage of NHK "News 7" ("News 7" is broadcasted every night in Japanese broadcast station NHK). It is 30 minutes long (about 50000 frames) for each day, which is broadcasted from Sep. 2001 to Feb. 2002.

3 Result of Identical Video Segment Detection

8164 pairs of identical video segments were found in the news archive. Several examples of identical video segments found on the news archive are already shown in Fig. 1. We noticed that each distribution of one of the same video segments differs quite significantly. Fig. 3 shows some example. Some segment distributes widely for the whole experiment period, while the others shrink in specific short periods.

In order to look into the result of the search, we also implement a news video browser using results of identical video segment detection. The interface of the browser is shown in Fig. 4. With the browser, we can display distribution of Identical Video Segment and we can also control display of distribution by masking some part of the archive.

First, we noticed that we can extract “important” topics by searching identical video segment detection. The browser shows the “distribution” of identical video segment detection. The “distribution” in this case is what is obtained by the following method as shown in Fig. 5. Each band in the figure represents each video stream of a day, and each stream is decomposed into a sequence of shots. Each arrow means that the shots indicated by the arrow are identical.

1. The result of identical video segment detection is a set of links between shots that shares the same frame image (a).
2. Extract shots that take part of the links (b).
3. Ignore link information (c).
4. Count number of shots for each broadcast date (d).

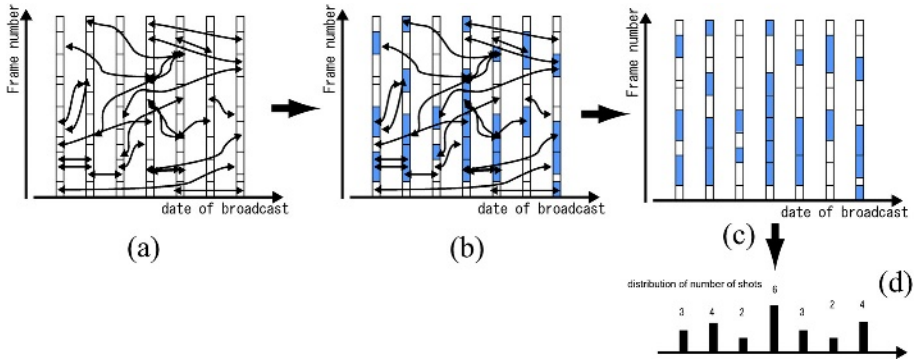


Fig. 5. Extracting distribution of identical video segments

The distribution indicates the degree that important topic existence because of the following reason. A news video is usually a sequence of several topics. If the exactly identical topic is broadcasted repeatedly (it is likely that exactly identical

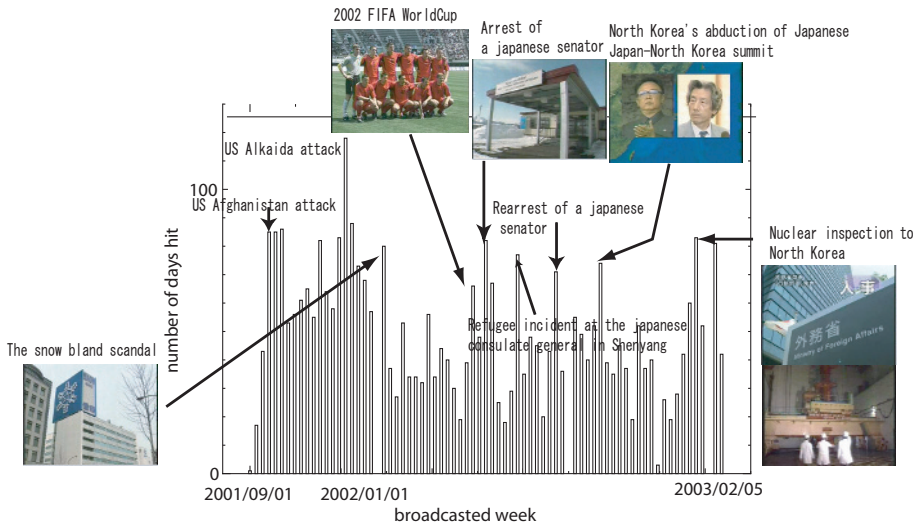


Fig. 6. Histogram of identical video segments and representative segments that make peaks

topics is broadcasted on morning and evening if no information comes between each broadcasting.), whole video of the topic is repeated. Two topics that have slight difference each other, it is highly probable that they share the same reference video. That is, the degree of share the same video segment correlates the degree of relation of those topics. By the way, it can be said that important topic is likely to be repeated with changing the content of the topic little by little. Of course, there may be important topics that are not repeated, but the amount of those topic is relatively little, so now we assume that repeated topics are “important” topics. Let’s look at the result, shown in Fig. 6. At each peaks of the distribution, you can find important topics of those days. It can be said that you can extract important topics for each time from video information only.

Next, we give an eye on time interval of identical video segments. With the browser, we can specify the two time period to select specific identical segments. Fig. 7 shows how we specify them. This is how we select specific video segments.

1. First, specify a time period in the broadcast time scale. Identical video segments that have their end in that period are selected. The Histogram of number of shots for each day is also shown.(Fig.7 (a))
2. Second, specify another time period to select video segments to display their representative images.(Fig.7 (b))

The browser displays identical video segment pairs only which has their one end in one period and the other end in the other period by the above process. So, we

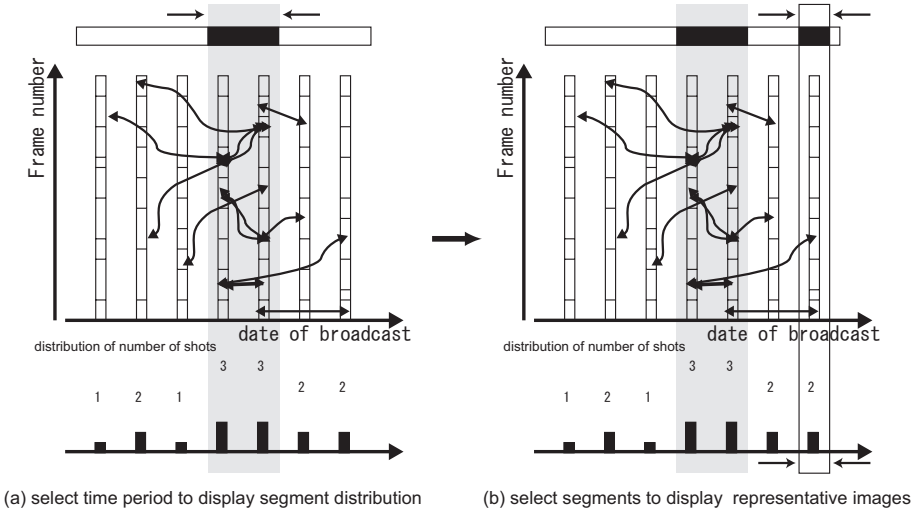


Fig. 7. Selecting segment pairs by broadcast time

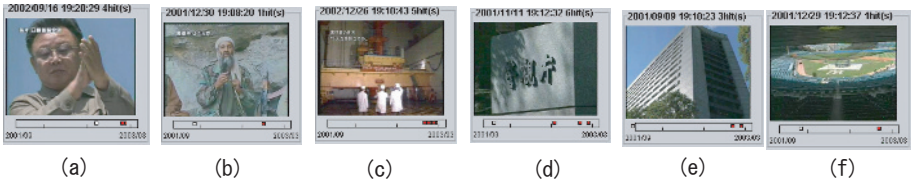


Fig. 8. Identical video segments with large time intervals



Fig. 9. Identical video segments with small time intervals

can control the time period of shown video segments. Now, we show two groups of identical video segments detected from the news archive for example. we found that identical video segments can be classified by time interval of appearance.

The first group of segments is found with long time interval (more than half year). Those segments are shown in Fig. 8. Most of this type of segments are very “rare” (Fig.8 (a)-(c)) or “symbolical” (Fig. 8 (d)-(f)). “Rare” in this case means that it is very difficult or impossible to get those segments again (ex. segments of an accident). Fig. 8 (a)-(c) are video segments that show Kim Jong Il, Usama

bin-Ladin and North Korean nuclear facility respectively, and it is very difficult to get their video information. “Symbolical” means that it is (almost) no use to get those segments again. Fig. 8 (d)-(f) are the doorplate of the building of Tokyo Metropolitan police, the building of the police and New York Yankee Stadium, respectively. They don’t change by time and each of them used as a symbol that represents some specific company or organization.

The second group of segments, which is shown in Fig. 9, is found with short time interval (several weeks at most). They are segments appeared in hot topics of each time.

4 Summary

In this paper, we applied identical video segment detection to actual news video archive. By browsing the result of identical video segment detection, people can know that important topics from distribution of identical video segments. We also observed that time intervals of appearances of identical video segments have some relation of types of video segment. For example, we can classify identical video segments by rareness or time-consistency of the segments, or hotness of the topics to which the segments belong. From above results, we believe that we can structurize video archives or extract useful information from them by searching identical video segments from the archives and analyzing the segments. We are still at the first point. We are going to do more precise analysis to identical video segments in large-scale video database.

References

1. Shin’ichi Satoh, “News video analysis based on identical shot detection”, Proc. of ICME, 2002.
2. S. S. Cheung and Avidah Zakhor, “Video Similarity Detection with Video Signature Clustering”, pp.649-652, Proc. of ICIP, 2001.
3. Norio Katayama and Shin’ichi Satoh, “Distinctiveness-Sensitive Nearest-Neighbor Search for Efficient Similarity Retrieval of Multimedia Information”, Proc. of ICDE, pp.493-502, 2001.
4. Fuminori Yamagishi, Shin’ichi Satoh, Takashi Hamada, Masao Sakauchi, “Identical Video Segment Detection For Large-Scale Broadcast Video Archive”, International Workshop on Content-Based Multimedia Indexing 2003.

Pseudo Relevance Feedback Based on Iterative Probabilistic One-Class SVMs in Web Image Retrieval*

Jingrui He¹, Mingjing Li², Zhiwei Li², Hong-Jiang Zhang², Hanghang Tong¹,
and Changshui Zhang³

¹ Automation Department, Tsinghua University, Beijing 100084, P.R.China
{hejingrui98, walkstar98}@mails.tsinghua.edu.cn

² Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P.R.China
{mjli, i-zli, hjzhang}@microsoft.com

³ Automation Department, Tsinghua University, Beijing 100084, P.R.China
zcs@tsinghua.edu.cn

Abstract. To improve the precision of top-ranked images returned by a web image search engine, we propose in this paper a novel pseudo relevance feedback method named iterative probabilistic one-class SVMs to re-rank the retrieved images. By assuming that most top-ranked images are relevant to the query, we iteratively train one-class SVMs, and convert the outputs to probabilities so as to combine the decision from different image representation. The effectiveness of our method is validated by systematic experiments even if the assumption is not well satisfied.

1 Introduction

With the ever-growing volume of digital images on the World Wide Web, the problem of how to effectively manage and index this huge image resource constantly draws people's research attention, and various web images search engines have been developed to address this issue, such as Google Image Search¹, AltaVista Image Search², AllTheWeb Picture Search³, etc.

When performing image retrieval using these text-based search engines, we can often find that some top-ranked images are irrelevant to the user's query concept. The problem may be attributed to the following reasons: the multiple meanings of words or phrases used to characterize the content of an image, misplacement of images in a totally irrelevant environment, etc. The removal of those top-ranked irrelevant images is highly desirable from the users' perspective.

One solution to this problem is to introduce relevance feedback into the retrieval process, which asks the user to mark the relevance of some retrieved images. However, in real applications, the user might be reluctant to provide

* This work was performed at Microsoft Research Asia.

¹ <http://www.google.com/imghp>

² <http://uk.altavista.com/image/>

³ <http://www.alltheweb.com>

any feedback information. Another possible solution is to re-rank the retrieved images before presenting them to the user, trying to put the relevant images in the head of the list while the irrelevant ones in the tail. It seems to be a promising way of solving the problem since no user intervention is required.

In the field of information retrieval, document re-ranking has long been studied to refine the retrieval result or to better organize the retrieved images [2][3][4]. Most of the techniques used in document re-ranking can be applied to web image retrieval in parallel. For example, Park et al [6] propose a hierarchical agglomerative clustering method to analyze the retrieved images; Yan et al [12] train SVMs whose positive training data are from the query examples, while negative training data are from negative pseudo relevance feedback; and Lin et al [5] propose a relevance model to calculate the relevance of each image, which is a probabilistic model that evaluates the relevance of the HTML document linking to the image. However, all of these re-ranking methods have drawbacks. For clustering based re-ranking methods, the number of clusters is hard to determine and it is quite doubtful that the constructed image clusters will be meaningful in all cases. For classification based approaches, positive training data is hard to obtain in the scenario of query by keyword. And the relevance model [5] depends on the documents returned by a text web search engine, which may be irrelevant with the retrieved images.

In this paper, we propose a novel pseudo relevance feedback method named iterative probabilistic one-class SVMs (IPOCS) to re-rank the retrieved images. Based on the assumption that most top-ranked images are relevant, given the images retrieved by a search engine, we iteratively train one-class SVMs (OCS) [9] for each image feature, convert their outputs to probabilities so as to combine the decision of various OCS, and re-rank the retrieved images according to their probabilities of being relevant. Systematic experiments demonstrate the effectiveness of our method, even if the assumption is not well satisfied.

The rest of the paper is organized as follows. In Sect.2, we present IPOCS in detail. To evaluate the proposed re-ranking method, we compare its performance with that of a recently developed manifold ranking algorithm [14] by systematic experiments in Sect.3. Finally, we conclude the paper in Sect.4.

2 Iterative Probabilistic One-Class SVMs

To better explain the method of IPOCS, in this section, we will first introduce OCS, followed by some discussion of its application in IPOCS; then we will discuss probabilistic outputs for OCS in order to combine the decision from different kinds of image features; finally, we will give the flowchart of the algorithm.

2.1 One-Class SVMs

The underlying principle of OCS is to estimate the minimum volume in the feature space that contains a constant fraction of the total probability of data distribution while keeping a large margin [8]. Among the several interpretations

of OCS [8][9], we present the most straight-forward one here. Consider training data $x_1, \dots, x_l \in X$, where $l \in \mathbb{N}$ is the number of observations, and X is the input space. Let Φ be a mapping $X \rightarrow F$ such that the dot product in the feature space F can be easily calculated by some kernel function: $k(x, y) = (\Phi(x) \cdot \Phi(y))$.

The basic idea of OCS is to find the smallest hyper-sphere to enclose most of the training data in the feature space F , which can be expressed as the following quadratic program:

$$\min_{R \in \mathbb{R}, \xi \in \mathbb{R}^l, c \in F} R^2 + \left(\sum_i \xi_i \right) / (vl) \quad (1)$$

$$\text{subject to } \|\Phi(x_i) - c\|^2 \leq R^2 + \xi_i, \xi_i \geq 0 \text{ for } i \in \{1, \dots, l\}$$

and the dual form:

$$\min_a \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) - \sum_i \alpha_i k(x_i, x_i) \quad (2)$$

$$\text{subject to } 0 \leq \alpha_i \leq 1/(vl), \sum_i \alpha_i = 1$$

with $c = \sum_i \Phi(x_i)$. The decision function takes the form:

$$f(x) = \text{sgn} \left(R^2 - \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) + 2 \sum_i \alpha_i k(x_i, x) - k(x, x) \right) \quad (3)$$

The parameter $v \in [0, 1]$ controls the tradeoff between the radius of the hyper-sphere and the number of observations that is enclosed in it. Furthermore, it has been proved that the following two statements hold [9]:

1. v is at least as large as the fraction of outliers;
2. v is no larger than the fraction of SVs.

Therefore, a large v ensures a large fraction of SVs and probably a small hyper-sphere; while a small v guarantees a small number of outliers and thus many observations within the hyper-sphere.

OCS has already been applied in content-based image retrieval [1], and the training data is provided by positive relevance feedback. In the context of web image retrieval, where no feedback information is available, we make a reasonable assumption that most top-ranked images are relevant, and use them as the training data. To remedy the possible errors incurred by this assumption, we design the following algorithm: given a certain kind of image representation, we first keep the top images for re-ranking. Then the first n ($n < m$) images are selected as pseudo positive examples, on which OCS is trained. Finally we re-rank the m images according to the outputs of OCS. The above operations are iterated until a certain criterion is satisfied.

The algorithm is justified as follows: the selection of the top m images for re-ranking is to ensure enough relevant images are included, and thus a high recall. On the other hand, the reason why only the first n images are used for training

OCS is that if many images are included in the training process, the presumed assumption might be violated, and the performance of OCS will deteriorate. The purpose for iterating between training and re-ranking is to progressively accumulate the relevant images in the head of the ranking list and the irrelevant ones in the tail.

In the training process of OCS, we take a large v . Thus many training observations will be support vectors, and their information will be fully utilized. The stopping criterion can be that the ranking order does not change after two consecutive re-ranking processes, or that the maximum iteration number is reached. In our implementation, the second criterion is taken to ensure the processing time is under control.

2.2 Probabilistic Outputs for OCS

In IPOCS, we construct OCS for each kind of image representation, the outputs of which must be combined for the overall decision. However, like two-class SVMs, OCS outputs uncalibrated values, thus we need to convert them to probabilities for the purpose of combination.

Inspired by the work presented in [7], after OCS is obtained, we train the parameters of an additional sigmoid function to map the outputs to probabilities. However, we use unlabeled data to train the sigmoid function, while in [7], the training data belongs to two classes. In the original algorithm [7], after SVMs is trained on observations of two classes, the posterior is modeled as follows:

$$P(y = 1|f) = 1/(1 + \exp(Af + B)) \quad (4)$$

where $f = f(x)$ is the uncalibrated output of SVMs for the observation x , $y \in \{-1, 1\}$ is the class label, and $P(y = 1|f)$ is the posterior probability that is a positive example given the output of SVMs. The parameters A and B are adapted to give the best probability outputs. As long as $A < 0$, (4) is monotonic, and large SVMs outputs correspond to large posterior probabilities.

The determination of A and B is based on maximum likelihood estimation, which can be transformed into the following optimization problem. Given a training set $\{(f_i, y_i)\}$, where f_i and y_i are the SVMs output and the class label of x_i , define $t_i = (y_i + 1)/2$ as the probability of x_i being a positive example. Note that since $y_i \in \{-1, 1\}$, $t_i \in \{0, 1\}$. Thus the optimization problem:

$$\min\left\{-\sum_i (t_i \log(p_i) + (1 - t_i) \log(1 - p_i))\right\} \quad (5)$$

where $p_i = 1/(1 + \exp(Af_i + B))$

To fit the sigmoid function, the algorithm in [7] uses a simple out-of-sample model: each observation in the training set is assigned with a small probability of opposite label in the out-of-sample data, i.e., $t_i \in [0, 1]$ instead of $t_i \in \{0, 1\}$. In our algorithm, this scheme is generalized to deal with unlabeled data. To speak

concretely, we establish a model for estimating the probability that an unlabeled image is a relevant one based on the original ranking result, i.e.

$$t_i = g(r_{x_i}) \quad (6)$$

where r_{x_i} is the ranking order of x_i , and $g(\cdot)$ is a function that maps each retrieved image to a real number in $[0, 1]$. In other words, the probability of an image being relevant is determined by its position in the original ranking list.

Generally speaking, $g(\cdot)$ should be a decreasing function, i.e., top-ranked images should have a large t_i , while bottom-ranked images should have a small one. In our current implementation, we take a simple form of $g(\cdot)$:

$$t_i = g(r_{x_i}) = 1/r_{x_i}^\beta \quad (7)$$

where β is a positive parameter that controls that decreasing rate of t_i as r_{x_i} increases. Presently, we set it to 1 for simplicity.

Once we obtain the posterior probabilities from the outputs of OCS constructed using each kind of image representation, we must combine the results to give the overall decision, in which several schemes may be taken. For example, we may average the probabilities, or select the largest one instead. In the context of web image retrieval, we prefer the second scheme to the first one, since we want to preserve the most confident decision of each OCS. Suppose that there are T kinds of image representation, let p^j denote the probabilistic output of OCS constructed using the j th image representation. The combination scheme can be expressed as follows:

$$p = \max\{p^1, \dots, p^T\} \quad (8)$$

2.3 The Flowchart of IPOCS

Based on the above discussion, we summarize the algorithm of IPOCS in Fig.1.

1. Perform similarity ranking, keeping the first m images for re-ranking;
2. Iterate for N_p times:
 - (a) Select the first n images as pseudo positive examples;
 - (b) for each kind of image representation:
 - Train OCS based on the pseudo positive examples;
 - Fit a sigmoid function to get the probabilistic outputs.
 - (c) Combine the outputs to get the overall probabilities, using (8);
 - (d) Re-rank the m retrieved images according to p .

Fig. 1. The flowchart of IPOCS

3 Experimental Results

In this section, we perform experiments to evaluate the performance of IPOCS, and compare it with manifold ranking (MR) [14], a recently developed algorithm for ranking data points which considers their global structure instead of pair-wise distances. We design two kinds of experiments in our evaluation: one is based on Corel images, and the other is based on the images returned by a prototype web image search engine *iFind* [13].

For both IPOCS and MR, we use the top $m = 100$ images for re-ranking, which are represented using color histogram [10] and wavelet features [11]. In IPOCS, we use these two kinds of image features to get two sets of probabilities, and fuse them to get the overall decision. The adopted kernel function in OCS is the RBF kernel, i.e., $k(x_i, x_j) = \exp[-\|x_i - x_j\|^2 / (2\sigma_p^2)]$. Thus there are four parameters needed to be set: n , v , σ_p and N_p . We conservatively set $n = 10$ to ensure that OCS will not be misled by many irrelevant images. Based on the discussion in subsection 3.1, we experimentally set $v = 0.99$, which achieves the best result among all the choices. The value of σ_p is empirically set to be 0.1. And N_p is set to 20 as a tradeoff between processing time and performance. In MR, the first n images are initially assigned with score 1, and the three parameters are set as follows: $\alpha = 0.99$, which is consistent with the experiments performed in [14]; $\sigma_m = \sigma_p$; and the iteration number N_m is set to 50, since we observe no improvement in performance with more iterations.

3.1 Experiments with Corel Images

We first form a general-purpose image database from which the initial retrieved images are to be selected. The database consists of 5,000 Corel images, which are made up of 50 image categories, each having 100 images of essentially the same topic. To simulate the top m images retrieved by a web image search engine, we first designate a certain category to contain all the relevant images, fix the ratio ra_m of relevant images in the m images, and randomly select images from the database according to ra_m . Then we vary the ratio ra_n of relevant images in the first n images to compare the two methods in different circumstances.

The adopted performance measure is precision. In this experiment, each of the categories is taken as the target, and the precision is averaged over all categories. The comparison results are illustrated in Fig.2.

From Fig.2, we can see that IPOCS can always significantly improve the retrieval result and outperform MR, no matter what value is taken for ra_m and ra_n . For example, when $ra_m = ra_n = 0.5$, P10 (precision within the top 10 images) is 82.5% using IPOCS, which improves the original precision (0.5) by 65.0%. While P10 is only 60.7% using MR, which improves the original precision by 21.4%. When $ra_m = 0.2$ and $ra_n = 0.5$, P10 is 76.9% using IPOCS, which improves the original precision by 53.8%. While P10 is 45.2% using MR, which even brings degradation to the original result. Note that when ra_n is small, the presumed assumption is not well satisfied, while IPOCS still significantly improves the ranking result.

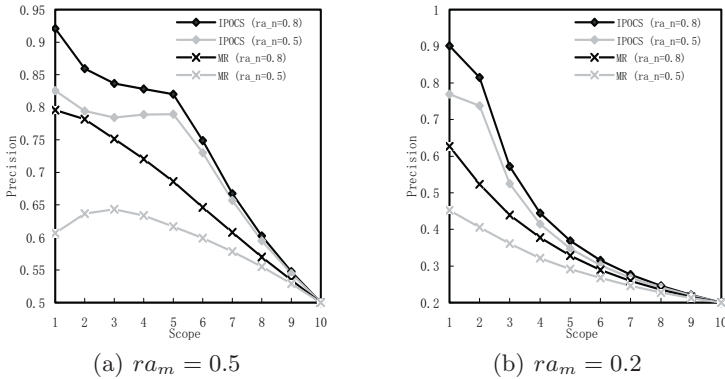


Fig. 2. Comparison of re-ranking results

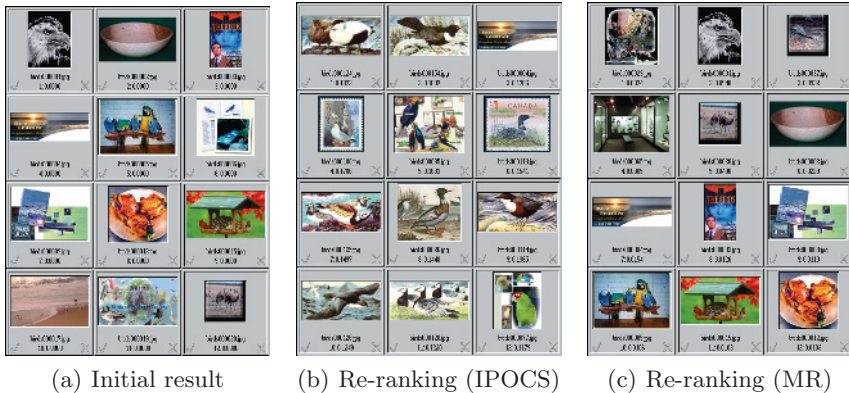


Fig. 3. Ranked images for the query “bird”

3.2 Experiments with Retrieved Images from *iFind*

iFind is a prototype web image search engine developed at Microsoft Research Asia. It is able to search 12 million web images based on text description of an image obtained from the web page where it is located. In this experiment, we resort to *iFind* to produce the initial retrieval result, given the query keywords, and make use of both IPOCS and MR to re-rank the images. In Fig.3, we compare the top 12 images using the two algorithms given the query keyword “bird”.

Obviously, the initial retrieval result Fig.3(a) produced by similarity ranking is far from satisfactory, since there are many irrelevant images due to inappropriate text description. Comparing both Fig.3(b) and 3(c) with the initial result, we can see that IPOCS greatly improves the performance, with the top 12 images all closely related to the query; while the improvement using MR is hard to tell.

4 Conclusion

In this paper, we have proposed a novel pseudo relevance feedback method based on IPOCS, which is used to re-rank images retrieved by a web image search engine. In the context where feedback information is not available, we make a reasonable assumption that most of the top-ranked images are relevant, and train OCS using these pseudo positive examples. To remedy the possible errors incurred by this assumption, we design an iterative algorithm to progressively refine the ranking result; furthermore, to combine the decision from different kinds of image representation, we train an additional sigmoid function which maps the outputs of OCS to probabilities. Experimental results demonstrate the effectiveness of the proposed method.

Acknowledgements. This work was supported by National High Technology Research and Development Program of China (863 Program) under contract No.2001AA114190.

References

- [1] Chen, Y., et al: One-class SVM for learning in image retrieval. Proc. ICIP (1999) 440-447
- [2] Hearst, M.A., et al: Reexamining the cluster hypothesis: Scatter/gather on retrieval results. Proc. SIGIR (1996) 76-84
- [3] Lee, K., et al: Document re-ranking model using clusters. KORTERM-TR-99-03 (1999)
- [4] Leuski, A.: Evaluating document clustering for interactive information retrieval. Proc. CIKM (2001)
- [5] Lin, W., et al: Web image retrieval re-ranking with relevance model. Proc. IEEE/WIC Int. Conf. on Web Intelligence (2003) 242-248
- [6] Park, G., et al: A ranking algorithm using dynamic clustering for content-based image retrieval. Proc. CIVR (2002) 328-337
- [7] Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, MIT Press (1999)
- [8] Ratsch, G., et al: Constructing boosting algorithm from SVMs: An application to one-class classification. *IEEE Trans. on PAMI* **24** (2002) 1184-1199
- [9] Scholkopf, B., et al: Estimating the support of a high-dimensional distribution. *Neural computation* **13** (2001) 1443-1471
- [10] Swain, M., et al: Color indexing. *Int. Journal of Computer Vision* **7** (1991) 11-32
- [11] Wang, J.Z., et al: Content-based image indexing and searching using Daubechies' wavelets. *Int. Journal of Digital Libraries*, **1** (1998) 311-328
- [12] Yan, R., et al: Multimedia search with pseudo-relevance feedback. Proc. Int. Conf. on Image and Video Retrieval (2003) 238-247
- [13] Zheng, C., et al: iFind: A web image search engine. SIGIR Demo (2001)
- [14] Zhou, D., et al: Ranking on data manifolds. 18th Annual Conf. on Neural Information Processing System (2003)

Robust Video Similarity Retrieval Using Temporal MIMB Moments*

Duan-Yu Chen and Suh-Yin Lee

Department of Computer Science and Information Engineering
National Chiao-Tung University, Taiwan
{dychen, sylee}@csie.nctu.edu.tw

Abstract. Due to the tremendous growth in the number of digital videos, the development of video retrieval algorithms that can perform efficient and effective retrieval task is indispensable. In this paper, we propose a high-level motion-pattern descriptor, temporal motion intensity of moving blobs (MIMB) moments, which exploits spatial and temporal features to characterize video sequences in a semantics-based manner. The Discrete Cosine Transform (DCT) is applied to convert the high-level features from the time domain to the frequency domain. The energy concentration property of DCT allows us to use only a few DCT coefficients to precisely capture the variations of moving blobs. Compared to the motion activity descriptors, RLD and SAH in MPEG-7, the proposed descriptor yield 41% and 20 % average performance gains over RLD and SAH, respectively. Having the efficient scheme for video representation, one can perform video retrieval in an accurate and efficient way.

1 Introduction

The tremendous growth in the number of digital videos has become the main driving force for developing automatic video retrieval techniques. Among different types of tools that can push the advancement of retrieval techniques, an efficient automatic content analyzer that can help execute correct browsing, searching and filtering of videos is a must. In order to achieve this goal, one has to make use of high-level semantic features to represent video contents. The need of representing high-level semantic features has motivated the emergence of MPEG-7, formally called the multi-media content description interface [1]. However, the methods that produce the specific features and the corresponding similarity measures represent the non-normative part of MPEG-7 and are still open for research and future innovation. Usually, the high-level semantic features of video sequences can be inferred from low-level features. The low-level features can be color distribution, texture composition, motion intensity and motion distribution. Among different types of features that can be extracted from a video, motion is considered as a very significant one due to its temporal nature. In the literature, Divakaran et al. [2] used a region-based histogram to

* This research is supported by NSC93-2219-E009-018, Taiwan

compute the spatial distribution of moving regions. The run-length descriptor in MPEG-7 [3] is used to reflect whether moving regions occurred in a frame. Aghbari et al. [4] proposed a motion-location based method to extract motion features from divided sub-fields. Peker et al. [5] calculated the average motion vectors of a P-frame and those of a video sequence to be the overall motion features. In addition to the above mentioned local motion features, Ngo et al. [6] and Tang et al. [7] proposed to use some global motion features to describe video content. In contrast to the motion-based features of individual frames, another group of researchers proposed to use spatio-temporal features between successive frames because these types of features are more abundant in the amount of information. Wang et al. [8] extracted features of color, edge and motion, and measured the similarity between temporal patterns using the method of dynamic programming. Lin et al. [9] characterized the temporal content variation in a shot using two descriptors – dominant color histograms of group of frames and spatial structure histograms of individual frames. Cheung and Zakhor [10] utilized the HSV color histogram to represent the key-frames of video clips and designed a video signature clustering algorithm for detecting similarities between videos. Dimitrova et al. [11] represented video segments by color super-histograms, which are used to compute color histograms for individual shots. There are several drawbacks associated with the key-frame based matching process. First, the features selected from key-frames usually suffer from the high dimensionality problem. Second, the features chosen from a key-frame is in fact local features. For a matching process that is targeting at measuring the similarity among a great number of video clips, the key-frame based matching method is not really feasible because the information used to characterize the relationships among consecutive frames is not taken into account. In order to overcome these drawbacks, we propose an moving blobs based motion pattern descriptor, which can exploit the spatio-temporal information of a video shot in the matching process. Basically, the proposed spatio-temporal features can support high-level semantic-based retrieval of videos in a very efficient manner. We make use of some spatio-temporal relationships among moving blobs and then use them to support the retrieval task. In the retrieval process, we use the DCT to reduce the dimensionality of the extracted high-dimensional feature. Using DCT, we can maintain the local topology of a high-dimensional feature. In addition, the energy concentration property of DCT allows us to use only a few DCT coefficients to represent the moving blobs and their variations. Therefore, the transformation can make an accurate and efficient retrieval process possible. The rest of the paper is organized as follows. Section 2 illustrates the methods used to characterize spatio-temporal features of video segments. Section 3 presents the experimental results. Section 4 draws conclusions.

2 Characterization of Video Segments

In this section, we shall describe how to characterize a video segment so that it can be used to perform efficient video retrieval. We shall describe how to detect

moving blobs in a video segment in Section 2.1 and then discuss how to compute the proposed spatial feature – MIMB moments in P-frames in Section 2.2. The proposed approach of characterizing spatio-temporal motion patterns in a video segment is shown in Section 2.3.

2.1 Detecting Moving Blobs in Compressed Domain

For computational efficiency, motion information in P-frames is used for the detection of moving blobs. In general, consecutive P-frames separated by two or three B-frames are still similar and would not vary too much. Therefore, it is reasonable to use P-frames as targets for detecting moving blobs. On the other hand, since the motion vectors estimated in MPEG videos is for the purpose of compression and thus may not be 100% correct, one has to remove the noisy part before they can be used. In our previous research [12], a cascaded filter that is composed of a Gaussian filter followed by a median filter is exploited for noise removal. The experimental results show that the precision is higher than 70% and the recall is higher than 80% and thus prove that the proposed spatial filter is effective to remove the noise in motion vector fields. To detect moving blobs in the filtered MVFs, macroblocks of similar MV magnitude and direction are clustered together by employing a region-growing method with an operator of 3x3 macroblocks.

2.2 Proposed Spatial Features of Moving Blobs – MIMB Moments

The motion intensity of moving blobs (MIMB) is a descriptor for describing sketch features in a frame that contain moving regions with motion intensity. Rather than directly employing the MIMB obtained in a P-frame, a temporal filter using Gaussian filter with temporal window size of 5 frames is exploited to smooth MIMBs. To represent the spatial feature of MIMBs in a compact meaningful form, the moment invariants of MIMBs are computed. The use of moments for image analysis and object representation was inspired by Hu[14]. According to Hu's Uniqueness Theorem, the moment set is uniquely determined by $MIMB(x,y)$ and conversely, $MIMB(x,y)$ is uniquely determined by the moment set. The central moment μ_{pq} computed from MIMB is defined by

$$\mu_{pq} = \sum_{y=0}^{C-1} \sum_{x=0}^{R-1} (x - \bar{x})^p (y - \bar{y})^q MIMB(x, y), \quad (1)$$

where $(p,q) = (0,2), (1,1), (2,0), (0,3), (1,2), (2,1), (3,0)$ and $C \times R$ denotes the total number of macroblocks in a P-frame. To select a meaningful subset of moment values that contain sufficient information to uniquely characterize the MIMBs, the seven moment invariants defined by Hu are employed and defined by

$$M_1 = \mu_{20} + \mu_{02} \quad (2)$$

$$M_2 = (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \quad (3)$$

$$M_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \tag{4}$$

$$M_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \tag{5}$$

$$M_5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12}) \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \tag{6}$$

$$M_6 = (\mu_{20} - \mu_{02}) \left[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \tag{7}$$

$$M_7 = (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12}) \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] + (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03}) \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \tag{8}$$

Therefore, a set of the moment invariants of MIMBs computed from central moments through order three, that are invariant to object scale, position, and orientation is exploited to characterize spatial features.

2.3 Representing Temporal Variations of Moving Blobs

In this section, we shall describe how to characterize the temporal variations among moving blobs exploiting the DCT. The algorithm that can be exploited to generate video sequence representation is as follows:

Video Sequence Representation Algorithm

Input: Consecutive P-frames $\{P1, P2, P3, \dots, PN\}$

Output: Sequences of representative truncated DCT coefficients $[X_{\Lambda,m}]$, where $\Lambda \in [1, \alpha]$.

Procedure:

1. For each P-frame Pi ,
 Detect moving blobs using a cascaded filter followed by using morphological operations.
2. For each P-frame Pi ,
 Compute Hu's seven moment invariants $\{M_{m,i}\}$ in the filtered MVF, where $m \in [1, 7]$.
3. Compute the transformed sequence $[X_{f,m}]$ using the Discrete Cosine Transform

$$X_{f,m} = C(f) \sum_{t=1}^N M_{m,t} \cos \left(\frac{(2t+1)f\pi}{2N} \right), \text{ where } f \in [1, N]$$

4. For m transformed sequences, $[X_{f,m}]$ of DCT coefficients,
 Truncate the number of DCT coefficients to α , which is composed of the DC coefficient and $(\alpha-1)$ AC coefficients to represent a transformed sequence.

5. Generate a feature vector $F(X_{A,1}, X_{A,2}, X_{A,3}, X_{A,4}, X_{A,5}, X_{A,6}, X_{A,7})$ for each video segment, where $A \in [1, \alpha]$.

For each P-frame, the spatial feature of moving blobs in P-frames is represented by Hu's seven moment invariants. In order to characterize the temporal variations of moving blobs within successive frames, DCT is exploited to transform the MIMB moments of the original video sequence into the frequency domain. The value of the MIMB $M_{m,i}$ in the i th P-frame is considered to be a signal in time i , and thus the corresponding MIMB $M_{m,i}$ in the N P-frames is regarded as a time signal $x_m = [M_{m,t}]$, where $T = 1, 2, 3, \dots, N$. The N -point DCT of a signal x_m is defined as a sequence $\mathbf{X} = [X_{f,m}]$, $F = 1, 2, 3, \dots, N$ as follows:

$$X_{f,m} = C(f) \sum_{t=1}^N M_{m,t} \cos\left(\frac{(2t+1)f\pi}{2N}\right), \quad (9)$$

$$C(0) = \sqrt{\frac{1}{N}} \text{ AND } C(f) = \sqrt{\frac{2}{N}}, f = 1, 2, \dots, N-1$$

where N is the number of P-frames and $m \in [1, 7]$. Eq.(9) indicates that a video sequence is represented by 7 sequences of DCT coefficients. It means that temporal variations among original objects in the successive P-frames are characterized by 7 sequences of DCT coefficients in frequency domain. It is well known that the first few low-frequency AC terms together with the DC term will suffice for the need. Therefore, for considering computation cost we only choose these terms to represent a video sequence instead of selecting all coefficients. However, to select an appropriate amount of AC coefficients is always a crucial issue. The experimental results imply that two DCT coefficients are enough for similarity measurement of video segments. This indicates the DC coefficient and the lowest-frequency AC coefficient will suffice.

3 Experimental Results and Discussions

3.1 Choice of Similarity Measure

The similarity measure is for computing the similarity between a feature vector of a query video shot and a feature vector of a target video shot. To choose a similarity measure, in statistics we prefer a distance that for each of the components takes the variability of that variable into account when determining its distance from the center. Components with high variability should receive less weight than components with low variability. Therefore, a Mahalanobis distance is used as a similarity measure, which is defined as

$$D(F^q, F^t) = \left(\sum_{k=1}^n \left| \frac{F_k^q - F_k^t}{\sigma_k} \right|^2 \right)^{1/2}, \quad (10)$$

where F_k^q and F_k^t denote the k th components of a query feature vector F^q and a target feature vector F^t , respectively and n denotes the dimension of a feature vector. σ_k denotes the standard deviation of the k th component for feature vectors in the testing dataset.

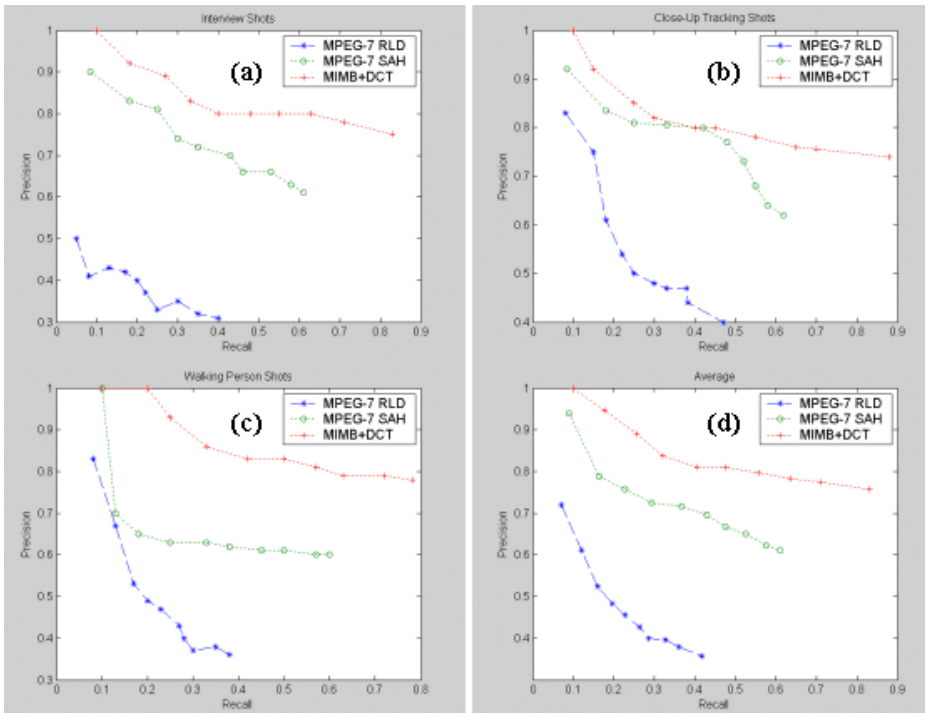


Fig. 1. Recall versus precision performance of the three shot classes (a) Interview Shots (b) Close-Up Tracking Shots (c) Walking Person Shots (d) Average

3.2 Evaluation of the Retrieval Performance

In order to show the effectiveness of the proposed method, we simulated the video sequence matching algorithm by using MPEG-7 testing dataset of Spanish News [13] which includes various programs such as news, sports, entertainment, education, etc and consists of 357 shots. The contents of the testing dataset mainly included the shots of an anchorperson, walking persons, football game, bicycle racing and interviews. The degree of strength of the motions in these shots ranged from low, medium to high, and the size of moving objects were classified as either small, medium or large. To evaluate the performance, precision and recall were used as the metrics to measure the performance of the proposed retrieval system. Recall and precision were defined as follows:

$$Recall = \frac{\|Retrieve(q) \cap Relevant(q)\|}{\|Relevant(q)\|}, Precision = \frac{\|Retrieve(q) \cap Relevant(q)\|}{\|Retrieve(q)\|} \quad (11)$$

where “ $Retrieve(q)$ ” means the retrieved video sequences that corresponded to a query sequence q ; “ $Relevant(q)$ ” denotes all video sequences in the database that were relevant to a query sequence q and $\|\cdot\|$ indicates the cardinality of the

set. *Recall* was defined as the ratio of the number of retrieved relevant video sequences to the total number of relevant video sequences in the video database, and *Precision* was defined as the ratio of the number of retrieved relevant video sequences to the total number of retrieved video sequences.

In the experiments, we used three kinds of shots to test the performance of our algorithms. Among these test videos, the shots covered in the Close-Up Tracking (CUT) and the Walking Person (WP) were with high degree of motion and medium degree motion, respectively. The Interview (IV) shots were with low degree of motion. Considering the sensitivity of the proposed descriptor to the blob size, the blob size in the test dataset ranges between small blobs of 2x2 macroblocks and large blobs of half or larger frame size. In the 30 most relevant shots corresponding to every query were selected out of 347 shots. In order to give a comparison, we also do the same experiments using the algorithms of motion-based run-length descriptor (RLD) and shot activity histogram (SAH) provided by MPEG-7. Fig. 1 shows the precision versus recall performance of the combination of RLD in MPEG-7 and the proposed MIMB+DCT descriptor. The proposed descriptor yielded 45% performance gain in the IV shots, 30% in the CUT shots, 34% in the WP shots, and 41% average over the RLD. Also, the proposed descriptor yielded 20% average performance gain in all the testing classes over the SAH descriptor of MPEG-7.

4 Conclusion

A novel framework of high-level video representation for video sequence matching has been developed in this work. The proposed framework has two special features: 1) the proposed temporal MIMB moments has exploited both spatial and temporal features of moving blobs and characterized video sequences in a high-level manner; 2) the dimensionality of feature space has been reduced using DCT while characterizing the temporal variations among moving blobs. Experimental results obtained using MPEG-7 testing dataset have demonstrated that a few DCT coefficients could suffice for representing a video sequence and also shown that the proposed motion-pattern descriptor was quite robust and efficient. Using this framework, one can perform video retrieval in an accurate and efficient way.

References

1. T. Sikora: The MPEG-7 Visual Standard for Content Description – An Overview. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, pp. 696 –702, June 2001.
2. A. Divakaran, K. Peker and H. Sun: A Region Based Descriptor for Spatial Distribution of Motion Activity for Compressed Video. Proc. International Conf. on Image Processing, Vol. 2, pp. 287-290, Sep. 2000.
3. S. Jeannin and A. Divakaran: MPEG-7 Visual Motion Descriptors. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, pp. 720-724, June 2001.

4. Z. Aghbari, K. Kaneko and A. Makinouchi: A Motion-Location Based Indexing Method for Retrieving MPEG Videos. Proc. 9th International Workshop on Database and Expert Systems Applications, pp. 102-107, Aug. 1998.
5. K. A. Peker, A. A. Alatan and A. N. Akansu: Low-Level Motion Activity Features for Semantic Characterization of Video. Proc. IEEE International Conference on Image Processing, Vol. 2, pp 801-804, Sep. 2000.
6. C. W. Ngo, T. C. Pong and H. J. Zhang: On Clustering and Retrieval of Video Shots. Proc. ACM Multimedia Conference, pp. 51-60, Ottawa, Canada, Oct. 2001.
7. Y. P. Tang, D. D. Saur, S. R. Kulkarni and P. J. Ramadge: Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 1, pp. 133-146, Feb. 2000.
8. R. Wang, M. R. Naphade, and T. S. Huang: Video Retrieval and Relevance Feedback in The Context of A Post-Integration Model. Proc. IEEE 4th Workshop on Multimedia Signal Processing, pp. 33-38, Oct. 2001.
9. T. Lin, C. W. Ngo, H. J. Zhang and Q. Y. Shi: Integrating Color and Spatial Features for Content-Based Video Retrieval. Proc. IEEE International Conf. on Image Processing, Vol. 2, pp. 592-595, Oct. 2001.
10. S. S. Cheung and A. Zakhori: Video Similarity Detection with Video Signature Clustering. Proc. IEEE International Conf. on Image Processing, Vol. 2, pp. 649-652, Sep. 2001.
11. L. Agnihotri and N. Dimitrova: Video Clustering Using SuperHistograms in Large Archives. Proc. 4th International Conference on Visual Information Systems, pp. 62-73, Lyon, France, Nov. 2000.
12. Ahmad, A.M.A., D. Y. Chen and S. Y. Lee: Robust Object Detection Using Cascade Filter in MPEG Videos. Proc. IEEE 5th International Symposium on Multimedia Software Engineering, pp. 196-203, Taichung, Taiwan, Dec 2003.
13. ISO/IEC JTC1/SC29/WG11/N2466: Licensing Agreement for the MPEG-7 Content Set. Atlantic City, USA, Oct. 1998.
14. M. Hu: Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, Vol. IT-8, pp. 179-187, Feb. 1962.

Complete Performance Graphs in Probabilistic Information Retrieval

N. Sebe¹, D.P. Huijsmans², Q. Tian³, and T. Gevers¹

¹ Faculty of Science, University of Amsterdam, The Netherlands

² Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

³ University of Texas at San Antonio, San Antonio, USA

Abstract. The performance of a Content-Based Image Retrieval (CBIR) system presented in the form of Precision-Recall or Precision-Scope graphs offers an incomplete overview of the system under study: the influence of the irrelevant items is obscured. In this paper, we propose a comprehensive and well normalized description of the ranking performance compared to the performance of an Ideal Retrieval System defined by ground-truth for a large number of predefined queries. We advocate normalization with respect to relevant class size and restriction to specific normalized scope values. We also propose new performance graphs for total recall studies in a range of embeddings.

1 Introduction

The performance characterization of content-based image and audio retrieval often borrows from performance figures developed over the past 30 years for probabilistic text retrieval. Landmarks in the text retrieval field are the books by Salton [1] and van Rijsbergen [2] as well as the proceedings of the annual ACM SIGIR and NIST TREC conferences.

In probabilistic text retrieval [2], TREC [3], and MPEG-7 descriptor performance evaluation [4], authors often go for single measure performance characterizations. These single measures are based on ranking results within a limited scope and in most cases they do take into account both the size of the relevant class and the effect of changing either the size or the nature of the embedding irrelevant items. By their nature these single measures have limited use, because their value will only have a meaning for standardized comparisons, where most of the retrieval parameters, such as the embedding, relevant class size, and scope are kept constant.

The results of performance measurements are often presented in the form of Precision-Recall and Precision-Scope graphs. Each of these standard performance graphs provides the user with incomplete information about how the Information Retrieval System will perform for various relevant class sizes and various embedding sizes. *Generality* (influence of the relevant fraction) as a system parameter hardly seems to play a role in performance analysis [5,6,7]. Although *generality* may be left out as a performance indicator when competing

methods are tested under constant generality conditions, it appears to be neglected even in cases where *generality* is widely varying (a wide range of relevant class sizes in one specific database is the most frequently encountered example).

The lack of generality information, in Precision-Recall and Precision-Scope graphs, makes it difficult to compare different sized IR Systems and to find out how the performance will degrade, when the irrelevant embedding is largely increased. Hence the performance of a scaled-up version of a prototype retrieval system cannot be predicted. The recent overview of [8] does not mention *generality* as one of the required parameters for performance evaluation. However, in [9] the authors convincingly show how the evaluation results depend on the particular content of the database. These considerations led us to re-evaluate the performance measurements for CBIR and the way these performance measures are visualized in graphs [10]. How can we make the performance measures for image queries on test databases more complete, so that results of specific studies cannot only be used to select the better method, but can also be used to make comparisons between different system sizes and different domains?

2 Performance Evaluation Elements

In a testing environment, the performance of the Retrieval System, in its selection of database items that are retrieved, should be compared to the equivalent situation where ground-truth has been constructed. An Ideal Information Retrieval System would mimic this ground-truth. Such an Ideal IR System would quickly present the user some or all of the relevant material and nothing more. The user would value this Ideal System as being either 100% effective or being without (0%) error. In [11], we referred to this Ideal System as the Total Recall Ideal System (TRIS). In practice, however, IR Systems are often far from ideal: generally the query results shown to the user (a finite list of retrieved elements) are incomplete (containing only some retrieved relevant class items) and polluted (with retrieved but irrelevant items).

We characterize a CBIR system using the following set of parameters:

$$\text{number of relevant items for a particular query} = \text{relevant class size} = c \quad (1)$$

$$\text{number of irrelevant items for a particular query} = \text{embedding size} = e \quad (2)$$

$$\text{ranking method} = m \quad (3)$$

$$\text{number of retrieved items from the top of the ranking list} = \text{scope} = s \quad (4)$$

$$\text{number of visible relevant items within scope} = v \quad (5)$$

$$\text{total number of items in the ranked database} = \text{database size} = c + e = d \quad (6)$$

In this set-up the class of relevant items is considered unordered and everything that precedes a particular ranking (like user feedback) is condensed into the *ranking method*. Performance is determined by the particular combination of the 4 free parameters, since the relevant outcome of a particular query, v , is a function of class size c , embedding size e , ranking method m , and scope s . However, in general, the average performance will be graphed for a number of ranking methods, to completely specify the retrieval system performance for a

Table 1. Categories and marginals for the contingency tables: P = Positive, N = Negative, FP = False Positive, FN = False Negative, TP = True Positive, TN = True Negative, R = Retrieved, NR = Not Retrieved, DB = Database size. In TRIS $v = s = c$ and $TP = P = R$.

v	$(c - v)$	c	TP	FN	P
$(s - v)$	$(d + v) - (c + s)$	e	FP	TN	N
s	$(d - s)$	d	R	NR	DB

ground checked set of queries. We also concentrate on retrieval settings where the embedding items vastly outnumber the relevant class items, $e \gg c$ and hence $d \approx e$:

$$v = v_m = f(c, d, s). \tag{7}$$

In our opinion a characterization of the Retrieval System performance should be based on the well-established decision support theory similar to the way decision tables or contingency tables are analyzed in [12]. From a quantitative decision-support methodology, our Query By Example (QBE) situation can be characterized for each ranking method by a series of decision tables [13] or, as they are also called, contingency tables [12]. A decision table for a ranking method represents a 2×2 matrix of (*relevant, irrelevant*) versus (*retrieved, not retrieved*) number of items for different choices of scope s , relevant class size c , and embedding e . It can also be seen as the database division according to the ground-truth versus its division according to Content-Based Information Retrieval at specific scope. The CBIR preferred choice of contingency table descriptors is given next to the Decision Support naming scheme in Table 1.

The performance or relevant outcome of the query, v from Eq. (7), can be normalized by division through either c , s , or d :

$$v/c = recall = r = f(1, d/c, s/c) = f(d/c, s/c) \tag{8}$$

$$v/s = precision = p = f(c/s, d/s, 1) = f(c/s, d/s) \tag{9}$$

$$v/d = f(c/d, 1, s/d) = f(c/d, s/d) \tag{10}$$

with $c/d = generality = g = expected\ random\ retrieval\ rate$.

Recall and *precision* are widely used in combination (Precision-Recall graph) to characterize retrieval performance usually giving rise to the well-known hyperbolic graphs from high *precision*, low *recall* towards low *precision*, high *recall* values. *Precision* and *recall* values are usually averaged over precision or recall bins without regard to class size, *scope*, or embedding conditions. That these are severe shortcomings can be seen from (8) and (9) where *recall* and *precision* outcomes are mutually dependent and may vary according to the embedding situation. To address these shortcomings, we propose to further normalize performance figures by restricting scopes to values that have a constant ratio with respect to the class sizes involved:

$$s_r = relevant\ scope = \frac{scope}{relevant\ class\ size} = \frac{s}{c} = \frac{r}{p} = a = constant \tag{11}$$

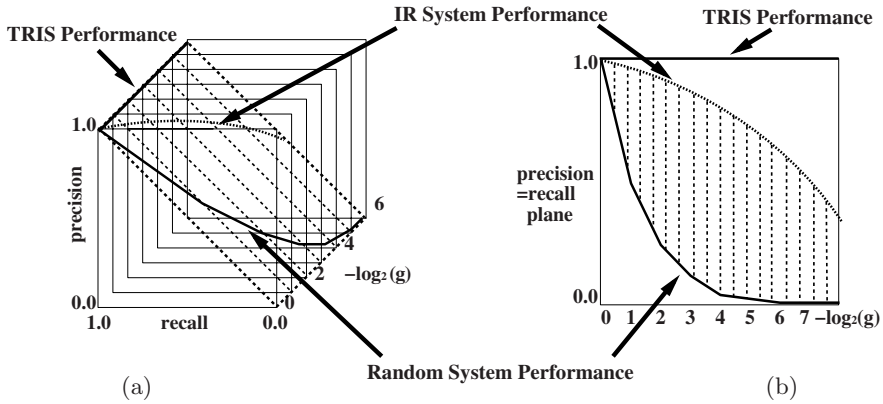


Fig. 1. (a) The 3D GReP Graph with the $p = r$ plane (with random and $s = c$ results for different generality values) holding the GRiP Graph; (b) The 2D GRiP Graph: $p = r$ values for scope size=relevant class size as a logarithmic function of generality.

With this relevant scope restriction, Eqs. (8) and (9) become:

$$r = f(1, d/c, ac/c) = f(1, d/c, a) = f(d/c) \tag{12}$$

$$p = r/a = f(c/ac, d/ac, 1) = f(1/a, d/ac, 1) = f(d/c). \tag{13}$$

This additional normalization of *scope* with respect to class size c means that the degrees of freedom for performance measures are further lowered from 2 to 1; only *recall* or *precision* values have to be graphed versus an embedding measure. Our preferred choice for the constant a in Eq. (11) is to set $a = 1$. With this setting one actually normalizes the whole Table 1 (now with $s = c$) by c , thus restricting ones view to what happens along the diagonal of the Precision-Recall Graph where $p = r$.

The only remaining dependency in this set-up (apart from the method employed) is on d/c . In Eq. (10) its inverse was defined as *generality* or the expected success-rate of a random retrieval method. Although generality g is a normalized measure, we will not graph it as such, because this would completely obscure the performance behavior for our case of interest, a range of $e \approx d \gg c$. Instead we propose to graph $p = r/a$ versus $-\log_2(g)$ to make the generality axis unbounded by giving equal space to each successive doubling of the embedding with respect to the relevant class size. We obtain thus the 3D Generality-Recall-Precision (GReP) graph (see Figure 1(a)).

The general 3D retrieval performance characterization, can be presented in 2D as a set of Precision-Recall graphs (for instance at integer logarithmic generality levels) to show how the p, r values decline due to successive halving of the relevant fraction. The two-dimensional graph, showing p, r values as a function of g (on a logarithmic scale), will be called the Generality-Recall=Precision Graph, GRiP Graph for short (see Figure 1(b)). For Total Recall studies, one could present several GRiP related graphs for planes in the GReP Graph, where $recall = n \cdot precision$: corresponding to the situation where the scope for retrieval

is a multiple of the relevant class size ($s_r = n$). We shall denote these Generality-Recall= n Precision Graphs as GR n P Graphs; obviously the GR 1 P Graph corresponds to the GR 1 P Graph.

In general, Precision-Recall graphs have been used as if the generality level would not matter and any p, r, g curve can be projected on a $g = \text{constant}$ plane of the three-dimensional performance space. However, our experiments reported in [14] show (at least for Narrow-Domain CBIR embeddings) that it does matter, and therefore Precision-Recall graphs should only be used to present performance evaluations when there is a more or less constant and clearly specified generality level. Only the Total Recall Ideal System (TRIS) as described for the PR graph is insensitive to generality by definition.

2.1 Scope Graphs Contained in P-R Graphs: Normalized Scope

Information about the effect of changing the *scope* on the measured *precision* and *recall* values can be made visible in the Precision-Recall graph by taking into account that possible *precision, recall* outcomes are restricted to lay on a line in the PR-graph radiating from the origin. This is due to the fact that the definitions of *precision* (Eq. (9)) and *recall* (Eq. (8)) have the same numerator v and are therefore not independent. The dependent pair of p, r values, and its relation to *scope*, becomes even more pronounced when *scope* is normalized with respect to the number of relevant items as defined by Eq. (11). Therefore, we present p, r values accompanied by their relevant scope line (radiating from the origin). So for each scope $s = a \cdot c$ with a an arbitrary positive number, $s_r = a$ and the p, r values are restricted to the line $p = r/a$. In Figure 2(a) we show several constant scope lines for retrieval of a relevant class of four additional relevant class members.

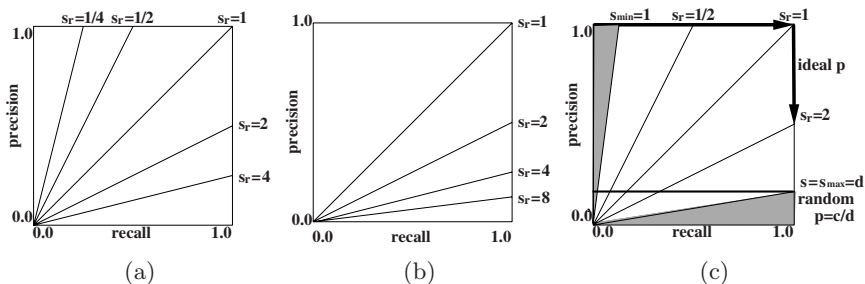


Fig. 2. (a) Lines along which p, r values are located at relevant class size=4 for several scopes; (b) Lines along which p, r values for retrieved relevant class size=1 (relevant class of 2, 1 used for query, max 1 for retrieval) are located; (c) p, r values for ideal retrieval are $1, r$ for $r < 1$; for scope size $>$ relevant class size, p drops slowly toward the random level, c/d .

With these relevant scope lines drawn in the Precision-Recall graph one understands much better what the p, r values mean. In the ideal case (see Figure 2(c)), *precision* p will run along $p = 1.0$ for *recall* $r \in [0.0, 1.0)$ and reach $p, r = 1.0, 1.0$ (the TRIS point) when *scope* equals relevant class size ($s = c$); for scopes greater than relevant class size, *precision* will slowly drop from $p = 1.0$ along $r = 1.0$ until the random level $p = c/d$ at $s = d$ is reached.

Also depending on relevant class size the region to the left of $p = r/c$ cannot be reached (solving the difficulty in PR-graphs for selecting a *precision* value for *recall* = 0.0) as well as the region below $p = dr/c$. This means that for the smallest relevant class of 2 members, where one of the relevant class members is used to locate its single partner, the complete upper-left half of the PR graph is out of reach (see Figure 2(b)).

Because the diagonal $s = c$ line presents the hardest case for a retrieval system (last chance of *precision* being max 1.0 and first chance of *recall* being max 1.0), and is the only line that covers all relevant class sizes (see Figure 2(b)), the best total recall system performance presentation would be the $p = r$ plane in the three-dimensional GRP Graph (Generality-Precision-Recall Graph).

2.2 Radial Averaging of Precision, Recall Values

For system performance one normally averages the discrete sets of *precision* and *recall* values from single queries by averaging *precision, recall* values without paying attention to the *generality* or *scope* values associated with those measurements. To compensate for the effect generality values have on the outcome of the averaging procedures, different ways of averaging are applied, like the micro- and macro-averaging used in text-retrieval [15]. In the critical review [16], the authors state with respect to averaging *precision* and *recall* values within the same database, that *precision* values should be averaged by using constant *scope* or cut-off values, rather than using constant *recall* values.

The fact stressed in Section 2.1, that p, r results have associated *generality* and relevant scope values, also has implications for the way average PR curves should be made up. Instead of averaging p, r values within recall or scope bins, one should average p, r values along constant relevant scope lines and only those that share a common *generality* value. When averaging for query results, obtained from a constant size test database, the restriction to averaging over outcomes of queries with constant relevant class sizes (constant generality value), will automatically result in identical micro- and macro-averages. The view expressed by [16] should therefore even be refined: the recipe, of averaging measured *precision, recall* values over their associated constant *scope* values only, should further be refined to our recipe of averaging p, r values over constant associated s_r, g values only.

An example of the way we determine an average p, r curve out of 2 individual curves with a shared generality value is given in Figure 3. The figure illustrates how averaging *recall* values in constant precision boxes (pbox-averaging) overestimates *precision* at low recall values, while underestimating it at high recall

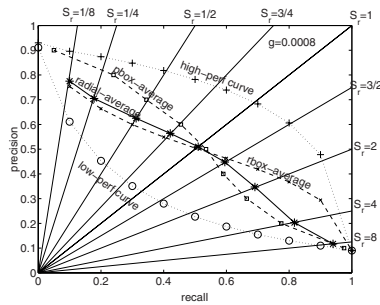


Fig. 3. Average PR-curves obtained from a low- and a high-performing PR-curve for 2 queries with class size 16 embedded in 21.094 images: the figure shows how large the difference can be between radial averaging compared to either precision-box (pbox) averaging or recall-box (rbox) averaging.

values; whereas averaging of *precision* values in constant recall boxes (rbox-averaging) underestimates *precision* at low recall while overestimating it at high recall values. In case of averaging discrete *precision, recall* values the errors introduced by not averaging radially (along constant relevant scope s_r) can be even more dramatic.

3 Laboratory Systems Versus Practical Systems

We have shown that for a complete performance evaluation, one has to carry out controlled retrieval tests, with queries for which ground-truth can provide the relevant class sizes. The performance is measured for various ranking methods, within a range of *scope* and *generality* values.

Since it is often too costly and labor intensive to construct the complete ground-truth for the queries used, we will indicate what could be done in terms of evaluation when knowledge about relevant class sizes c , and as a result *recall* and *generality* values, are missing.

First, let us make a distinction between Laboratory and Practical CBIR systems. We propose to reserve the name Laboratory CBIR system for those performance studies where complete ground-truth has become available. For these systems a complete performance evaluation, in the form of Generality-Recall-Precision Graphs, for a set of test queries and for a number of competing ranking methods can be obtained.

Any CBIR retrieval study that lacks complete ground-truth will be called a Practical system study. In Practical system evaluation one normally has a set of queries and a database of known size d . Because ground-truth is missing, relevant class size c is unknown. The only two free controls of the experimenters are the scope s and the ranking method m . Relevance judgments have to be given within the scopes used to determine the number of relevant answers. Of the three Laboratory system evaluation parameters *precision, recall*, and *generality* only

precision = v/s is accurately known. For *recall* due to knowing v but not c only a lower bound $v/(d-s+v)$ is known. For *generality* only a lower bound $g = v/d$ is known. In general, for practical studies, one characterizes the performance as Precision-Scope Graphs or one uses single measures obtained from the weighted ranks of the relevant items within scope.

The problem with any Practical system study is that one cannot interpret the results in terms of "expected completeness" (recall), and the results are therefore only useful in terms of economic value of the system: how many items will I have to inspect extra, to obtain an extra relevant item? Actually, with some extra effort the analysis of a Practical system can be enhanced to that of an estimated Laboratory system, by using the fact that *generality* in terms of relevant fraction is identical to the expected *precision* (see Eq. (10)) when using a random ranking method. Experimenters that have access to the ranking mechanism of a retrieval system can thus obtain estimates for generality g , and hence estimates for relevant class size c and recall r to complete their performance evaluation. The extra effort required would be the making of relevance judgments for a series of randomly ranked items within some long enough scope for each query.

4 Conclusions

We surveyed how the role of embeddings in Content-Based Image Retrieval performance graphs is taken care of and found it to be lacking. This can be overcome by adding a generality component. We also noted that one is not aware of the scope information present in a Precision-Recall Graph and the lack of comparison with random performance. The present practice of averaging *precision*, *recall* values in recall or precision boxes, conflicts with the way *precision* and *recall* are dependently defined.

We conclude that, Precision-Recall Graphs can only be used when plotting *precision*, *recall* values obtained under a common, mentioned, *generality* value which coincides with the random performance level. Therefore, to complete performance space we extended the traditional 2D Precision-Recall graph to the 3D GRP Graph (Generality-Recall-Precision Graph) by adding a logarithmic generality dimension. Moreover, due to the dependency of *precision* and *recall*, their combined values can only lay on a line in the PR Graph determined by the *scope* used to obtain their values. Scopes, therefore, can be shown in the PR Graph as a set of radiating lines. A normalized view on scope, relevant scope, makes the intuitive notion of scope much simpler. Also, averaging *precision*, *recall* values should be done along constant relevant scope lines, and only for those p, r values that have the same *generality* value.

References

1. Salton, G.: The SMART retrieval system. Prentice Hall (1971)
2. van Rijsbergen, C.: Information Retrieval. Butterworths, London (1979)

3. Voorhees, E.M., Harman, D.: Proc. TREC. (1999)
4. MPEG-7. Special Issue of IEEE Trans. Circuits and Systems for Video Technology **11** (2001)
5. Porkaew, K., Chakrabarti, K., Mehrotra, S.: Query refinement for multimedia similarity retrieval in MARS. In: ACM Multimedia. (1999) 235–238
6. Vasconcelos, N., Lippman, A.: A probabilistic architecture for content-based image retrieval. In: CVPR. (2000) 216–221
7. Baumgarten, C.: A probabilistic solution to the selection and fusion problem in distributed information retrieval. In: SIGIR. (1999) 246–253
8. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Performance evaluation in content-base image retrieval: Overview and proposals. Pattern Recog. Letters **22** (2001) 593–601
9. Müller, H., Marchand-Maillet, S., Pun, T.: The truth about Corel - Evaluation in image retrieval. In: CIVR 2002. (2002) 38–49
10. Huijsmans, D., Sebe, N.: Extended performance graphs for cluster retrieval. In: CVPR. (2001) 26–31
11. Huijsman, D.P., Sebe, N.: How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. IEEE Trans. on PAMI, to appear (2004)
12. Gokhale, D., Kullback, S.: The Information in Contingency Tables. M. Dekker (1978)
13. van Bommel, J.H., Musen, M.A.: Handbook of Medical Informatics. Springer (1997)
14. Huijsmans, D., Sebe, N.: Content-based indexing performance: A class size normalized precision, recall, generality evaluation. In: ICIP. (2003) 733–736
15. Tague-Sutcliffe, J.: The pragmatics of information retrieval experimentation, revisited. Information Processing and Management **28** (1992) 467–490
16. Raghavan, V., Bollmann, P., Jung, G.: A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Information Systems **7** (1989) 205–229

A New MPEG-7 Standard: Perceptual 3-D Shape Descriptor

Duck Hoon Kim¹, In Kyu Park², Il Dong Yun³, and Sang Uk Lee¹

¹ School of Electrical Engineering and Computer Science, Seoul National University,
SEOUL 151-742, KOREA

ducks@diehard.snu.ac.kr, sanguk@ipl.snu.ac.kr

² School of Information and Communication Engineering, INHA University,
INCHEON 402-751, KOREA

pik@ieee.org

³ School of Electronic and Information Engineering, Hankuk University of Foreign Studies,
YONGIN 449-712, KOREA

yun@hufs.ac.kr

Abstract. In this paper, we introduce the perceptual 3-D shape descriptor (P3DS) for 3-D object, which has been proposed and adopted as an MPEG-7 standard recently. The descriptor is used as an abstract representation of 3-D object in content-based shape retrieval system. Unlike the conventional descriptors, the P3DS descriptor supports the functionalities like *Query by Sketch* and *Query by Editing* which are very useful in real retrieval system. Given a couple of the descriptors, a matching technique is also developed to measure the similarity between them. High retrieval score has been observed when the developed descriptor and the matching technique are used in the retrieval system with the MPEG test database.

1 Introduction

As the applications of 3-D computer graphics become more popular in human life, the amount of 3-D object we need to handle increases day by day. In order to reuse the existing 3-D objects, they should be stored in an archive and retrieved again for further use when necessary. Therefore, efficient methods for managing 3-D objects are quite necessary. In this context, multimedia industry has been developing an international standard for multimedia contents description and retrieval, known as MPEG-7. Recently, MPEG-7 community has witnessed the necessity of a description scheme for 3-D object. In this paper, we introduce the perceptual 3-D shape descriptor (P3DS) for 3-D object, which has been proposed and adopted as an MPEG-7 standard recently [1,2,3,4,5,6,7]. Note that an amendment of the P3DS descriptor has been in progress and currently in Working Draft stage [7]. The P3DS descriptor provides the abstract and compact representation of the conventional objects in polygonal mesh, B-spline surface, and voxel representation. The topological structure of 3-D object is preserved in the P3DS descriptor, providing an interactive query design in content-based retrieval system. Based on the perceptual description as a building block, an efficient attributed relational graph matching method, *i.e.* Double Earth Mover's Distance, is proposed and optimized in the shape retrieval system.

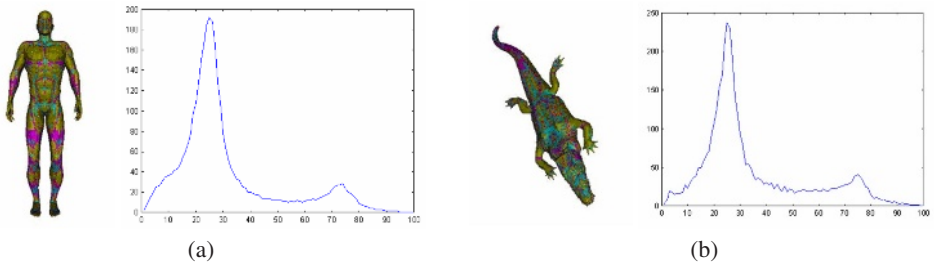


Fig. 1. The limitation of the discriminative ability in the Shape3D descriptor: (a) the android and its shape spectrum and (b) the crocodile and its shape spectrum.

2 Previous 3-D Shape Descriptors

In MPEG-7, the Shape3D descriptor has been developed and adopted as an international standard for description and browsing of 3-D VRML (Virtual Reality Markup Language) database [8]. The Shape3D descriptor implies the shape spectrum of 3-D mesh object, which is the histogram of the shape indices calculated over the entire mesh. It yields high retrieval scores for the ground truth data set [9]. However, the Shape3D descriptor has a few drawbacks. First, the shape spectrum, *i.e.* the histogram of the shape indices, represents only local geometries of the 3-D surfaces so loses the spatial information. Hence, the discriminative ability may be limited. As shown in Fig. 1, the android and the crocodile have different topological structures with different shape. However, two objects have very similar shape spectrums, *i.e.* the distance between two objects is only 0.025981. As a consequence, the crocodile is probably retrieved when the android is selected as query, which is an undesirable result. Next, the similarity measure of two shape spectrums can not evaluate correct distance since neighboring bins are not considered when there is no match between the exact corresponding bins.

In academia, a lot of research groups have developed various 3-D shape descriptors. Especially, people in the research group of Princeton University have been showing remarkable activities in 3-D shape retrieval [10,11,12,13]. More specifically, they have developed various descriptors [10,11,12], proposed the benchmark framework for 3-D shape retrieval, and performed the experiments for the comparison between existing 3-D shape descriptors [13]. However, their 3-D shape descriptors are mainly based on the distribution of vertices or faces so that it is difficult to provide the user-friendly querying interface such as *Query by Sketch* and *Query by Editing* which will be described in Section 3.3.

3 Perceptual 3-D Shape (P3DS) Descriptor

In order to overcome the drawbacks of the previous descriptors, we develop the perceptual description based on the part-based representation of a given object. The part-based representation is expressed by means of an attributed relational graph (ARG) which can be easily converted into the P3DS descriptor. More specifically, the P3DS descriptor is

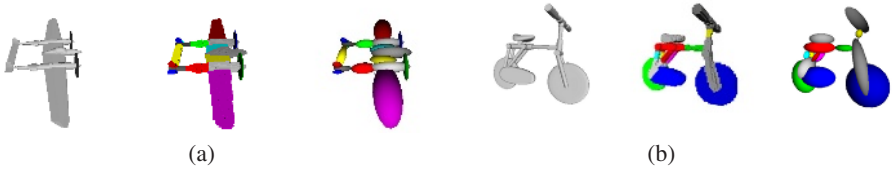


Fig. 2. Part-based and ARG representations of different 3-D objects. From left to right in (a) and (b), original mesh object, part-based representation, and ARG are shown sequentially.

designed to represent and identify 3-D objects based on the part-based simplified representation using ellipsoidal blobs. Actually, it is based on the assumption that the part-based representation and the actual shape are coherent with human visual perception. In this context, if volume-based decomposition is adopted as the part-based representation, the object is assumed to have its own volume, *e.g.* the object is composed of one or more closed mesh surface. On the other hand, in the case of mesh-based decomposition, the object is assumed to be manifold, *i.e.* an edge is shared by two triangles, unless it belongs to the boundary. For the fast processing and better result, it is recommended to use manifold mesh object with no hole. As expected, if the encoder does not produce the part-based representation properly, the retrieval performance would not be good. In this section, the P3DS descriptor is introduced with an example of its former self, *i.e.* ARG, and with the formal syntax in MPEG-7 standard. Moreover, new functionalities of the P3DS descriptor are addressed as well.

3.1 Node and Edge Features in the ARG

The perceptual description of a given 3-D object is generated from the part-based representation. It has the form of an ARG, composed of a few nodes and edges. A few examples of the part-based representation and ARG are shown in Fig. 2, where the morphological voxel-based decomposition [14] is adopted as the part-based representation.

A node represents a meaningful part of the object with unary attributes, while an edge implies binary relations between nodes. In the descriptor, there are 4 unary attributes and 3 binary relations which are derived from the geometric relation between the principal axes of the connected nodes. In detail, a node is represented by an ellipsoid parameterized by volume V , convexity C , and two eccentricity values E_1 and E_2 . More specifically, the convexity is defined as the ratio of the volume in a node to that in its convex hull, and the eccentricity is composed of two coefficients, $E_1 = \sqrt{1 - c^2/a^2}$ and $E_2 = \sqrt{1 - c^2/b^2}$, where a , b , and c ($a \geq b \geq c$) are the maximum ranges along 1st, 2nd, and 3rd principal axes, respectively. Edge features, *i.e.* binary relations between two nodes, are extracted from the geometric relation between two ellipsoids, in which the distance between centers of connected ellipsoids and two angles are used. The first angle is between first principal axes of connected ellipsoids and the second one is between second principal axes of them.

Perceptual3DShape {	Number of bits	Mnemonics
numberOfNodes	8	uimsbf
BitsPerAttribute	4	uimsbf
for (i = 0 ; i < (numberOfNodes ² - numberOfNodes) / 2 ; i++) {		
IsAdjacent[i]	1	bslbf
}		
for (i = 0 ; i < numberOfNodes ; i++) {		
Volume[i]	BitsPerAttribute	uimsbf
Center_X[i]	BitsPerAttribute	uimsbf
Center_Y[i]	BitsPerAttribute	uimsbf
Center_Z[i]	BitsPerAttribute	uimsbf
Transform_1[i]	BitsPerAttribute	uimsbf
Transform_2[i]	BitsPerAttribute	uimsbf
Transform_3[i]	BitsPerAttribute	uimsbf
Transform_4[i]	BitsPerAttribute	uimsbf
Transform_5[i]	BitsPerAttribute	uimsbf
Transform_6[i]	BitsPerAttribute	uimsbf
Variance_X[i]	BitsPerAttribute	uimsbf
Variance_Y[i]	BitsPerAttribute	uimsbf
Variance_Z[i]	BitsPerAttribute	uimsbf
Convexity[i]	BitsPerAttribute	uimsbf
}		
}		

Fig. 3. The binary representation syntax of the P3DS descriptor.

3.2 Binary Representation of the P3DS Descriptor

In order to preserve the topological shape of the object and to adopt the user-friendly querying interface in the retrieval system, the actual descriptor utilizes slightly different attributes which can provide the shape and relation of ellipsoidal blobs more intuitively. In detail, the P3DS descriptor contains three node attributes, such as Volume, Variance, and Convexity, which can be converted easily into the 4 unary attributes. Next, it contains two edge attributes, such as Center and Transform, from which the 3 binary relations can be computed as well. Fig. 3 shows the Binary Representation Syntax of the P3DS descriptor. Actually, the size of the descriptor is very compact. For an example, when an ARG has 5 nodes, the P3DS descriptor, of which the attributes are quantized in 8-bit (default), can be represented by only 582 bits. Note that Volume, Center, Variance and Convexity are normalized in the interval [0, 1], while Transform is normalized in the interval [-1,1].

3.3 New Functionalities for 3-D Shape Retrieval

Using the P3DS descriptor, we believe that it would produce a few new functionalities of 3-D shape retrieval, which cannot be provided by the previous descriptors. Note that the most important advantage of the P3DS descriptor is that there is the exact coincidence between the perceptual description in the P3DS descriptor and human cognitive description. Therefore, when we see the descriptor, we can understand the topological shape of

3-D object and also expect the proper retrieval result. For example, if the descriptor has some information of 5 parts in which four of them are connected to a main body, then we can perceive that the descriptor represents animal-like shape and we would expect the retrieval results of such animals. The human readability of the descriptor enables us to design new kinds of querying strategy as follows: Since the part-based representation is very easy to build and edit, users can make a simple shape consisting of ellipsoidal blobs and their connections. The built object is easily translated to the P3DS descriptor and put into the search engine, *i.e.* *Query by Sketch*. Similarly, users can modify the existing descriptor in an interactive way and then they can try a new retrieval to get different results, *i.e.* *Query by Editing*.

4 Shape Matching Using the P3DS Descriptor

The similarity measurement of two P3DS descriptors is the most basic procedure since the input query is compared with the model in the database one by one using the P3DS descriptor and then the models with high similarity are browsed. The one-to-one comparison of two P3DS descriptors consists of three steps: (1) Forming ARG's from the descriptors, (2) Defining the Volume as weight in each node, (3) Comparing the ARG's using the Double EMD (Earth Mover's Distance [15]) method. During the Double EMD procedure, a distance matrix is first generated between query nodes of a query ARG and model nodes of a model ARG (Inner EMD), followed by measuring the similarity by calculating the amount of work required to move the weights from the query nodes to the model nodes using the conventional EMD method based on the distance matrix (Outer EMD).

The Inner EMD is defined as follows. First, the unary-distance matrix is constructed by computing the distance between the unary attribute vectors of the query and model ARG's. Next, to define the difference of binary relations between nodes of query and model ARG's, *i.e.* N_q and N_m in Fig. 4(a), as one element of the binary-distance matrix, the binary relations are utilized to form the axes of the vector space. More specifically, the coordinate system in Fig. 4(a) consists of three axes, such as one distance and two angles defined in Section 3.1. Then N_q and its connected nodes form a set of points in the vector space, while N_m and its connected nodes form another set of points. In this context, N_q and N_m are located at the origin of the vector space. Finally, an imaginary point, *i.e.* empty circles in Fig. 4(a), of which the distance from any point is equal to d in Fig. 4(b), is provided to make the sum of the weights of N_q and its connected nodes and that of N_m and its connected nodes be equal. Note that the imaginary point can penalize the weight transition to other unconnected nodes. As shown in Fig. 4(b), a distance matrix with respect to N_q and N_m is constructed from the Euclidean distances between nodes of query and model ARG's. Then the conventional EMD based on this distance matrix yields one element of the binary-distance matrix. In this way, each element of the binary-distance matrix can be computed. Finally, the final distance matrix is computed by adding the unary-distance matrix and the binary-distance matrix.

In the Outer EMD, the dissimilarity between the query and model ARG's is measured by calculating the amount of work required to move the weights from the query nodes

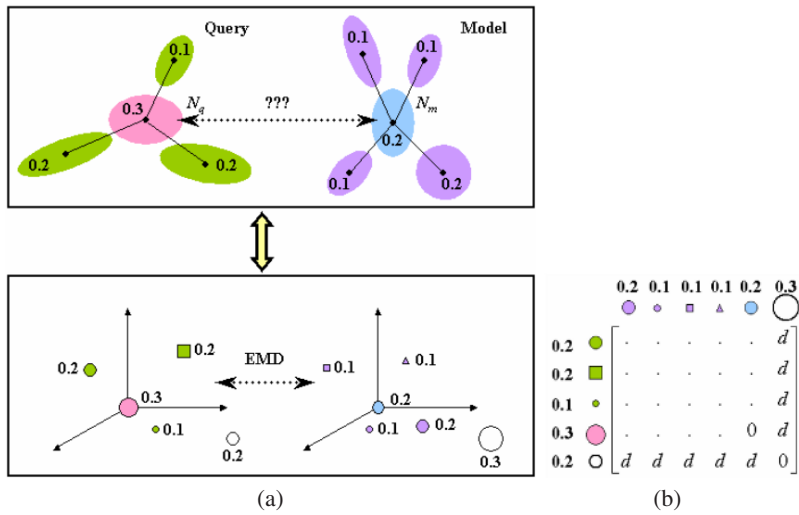


Fig. 4. An example of Inner EMD: (a) vector space representation for computing the Inner EMD and (b) a distance matrix in the procedure of calculating one element of the binary-distance matrix considering N_q and N_m .

to the model nodes based on the final distance matrix. In other words, a total amount of work for all of the nodes refers to the dissimilarity between the two ARG's.

5 Experimental Results

In order to evaluate the performance of the developed 3-D shape retrieval system using the P3DS descriptor, intensive experiments are carried out on the MPEG-7 official 3-D VRML database including 3,900 objects. The objects in the database are classified into 102 categories in a hierarchical structure. For the query dataset, we also use the MPEG-7 query dataset composed of 336 objects in 10 leaf categories from 8 broad categories. Note that MPEG's basic rule to select the leaf categories is that the objects should have perceptually meaningful ARG structure. In other words, the object itself and its ARG should be similar enough based on the human visual ability in object recognition. In the experiment, we show the retrieval performance and the comparison with the Shape3D descriptor. Note that d in Section 4 is set to 1 experimentally. As shown in Table 1, it is observed that the performance of the P3DS descriptor is quite good for both Bull's Eye Performance (BEP) and Average Normalized Modified Retrieval Rate (ANMRR) [8], also significantly better than the Shape3D descriptor. Note that higher BEP score and lower ANMRR score imply better performance. Moreover, the complexity of the retrieval is quite endurable. The retrieval time depends on the complexity of the object structure. For most cases, it would take less than 5 seconds on a PC with Pentium IV CPU. Currently, we use simple one-to-one matching for the whole objects in the database. We believe it is necessary to develop an algorithm for fast indexing which fully utilizes the hierarchical database structure.

Table 1. Retrieval performance and comparison with the Shape3D descriptor (Higher BEP and lower ANMRR scores show better performance)

Category		P3DS descriptor		Shape3D descriptor	
		BEP	ANMRR	BEP	ANMRR
<i>Aircraft / multi_fuselages / 3_bodies</i>	30	0.63667	0.27326	0.32111	0.51363
<i>Animal / arthropod / with_wings / bee</i>	30	1.00000	0.00000	0.55556	0.31395
<i>Animal / humanoid / sitting</i>	60	0.80972	0.17468	0.55556	0.35791
<i>Automobile / tank / equipvaried</i>	30	1.00000	0.00007	0.53222	0.32587
<i>Furniture / chair / 4_legged</i>	30	0.58556	0.32306	0.34333	0.57812
<i>Furniture / chair / with_a_post</i>	30	0.68333	0.24401	0.25000	0.60753
<i>Letter / O</i>	15	0.58667	0.29977	0.32000	0.51549
<i>Plant / flower / 20_petaled</i>	30	0.66111	0.23979	0.49222	0.38677
<i>Ship / single_mast / romanship</i>	30	0.93333	0.05263	0.33778	0.49117
<i>Simplex cellular_phone</i>	51	0.85621	0.11558	0.36140	0.47758
Total	336	0.79182	0.16326	0.42122	0.44665

Table 2. Influence of the quantization steps in the P3DS descriptor for BEP score (Higher BEP score shows better performance)

Category		BEP				
		8 bits	7 bits	6 bits	5 bits	4 bits
<i>Aircraft / multi_fuselages / 3_bodies</i>	30	0.63667	0.63222	0.64444	0.66444	0.55556
<i>Animal / arthropod / with_wings / bee</i>	30	1.00000	1.00000	1.00000	1.00000	1.00000
<i>Animal / humanoid / sitting</i>	60	0.80833	0.80861	0.80000	0.74861	0.60528
<i>Automobile / tank / equipvaried</i>	30	1.00000	1.00000	1.00000	1.00000	0.98778
<i>Furniture / chair / 4_legged</i>	30	0.59000	0.59556	0.59556	0.60667	0.49667
<i>Furniture / chair / with_a_post</i>	30	0.68000	0.67444	0.68222	0.67111	0.68444
<i>Letter / O</i>	15	0.59111	0.59111	0.58667	0.52889	0.43556
<i>Plant / flower / 20_petaled</i>	30	0.66000	0.66778	0.67333	0.62333	0.56667
<i>Ship / single_mast / romanship</i>	30	0.93222	0.93111	0.93111	0.93222	0.92333
<i>Simplex cellular_phone</i>	51	0.85659	0.85429	0.85275	0.83045	0.80238
Total	336	0.79132	0.79162	0.79194	0.77422	0.71490

Actually, the P3DS descriptor consists of an array of floating point (32bits) attributes. In this context, the quantization of the attributes with lower number of bits should be considered in order to achieve more compact representation. In our implementation, we perform an uniform quantization with a number of steps 2^b , where b denotes the number of quantization bits. Table 2 presents the influence of the quantization bits to the overall retrieval performance. Note that 8 bits uniform quantization produces very little degradation of the performance compared with the results without quantization. Moreover, lower bits representation up to 5 bits still produces quite little degradation, while the performance goes down much at 4 bits.

6 Conclusion

In this paper, we introduced the Perceptual 3-D Shape (P3DS) descriptor and presented the various experimental results not only in shape retrieval but also in the effect of the attribute quantization. In MPEG-7 community, the P3DS descriptor has been proposed and adopted as a new standard for 3-D shape description and currently in Working Draft stage. Our future work includes the development of practical retrieval system which can be used in real application. The combination of context-based (keyword-based) and content-based retrieval system should be taken into account. Furthermore, some scheme like relevance feedback is quite necessary in order to increase the subjective performance to real user.

References

1. Kim, D.H., Yun, I.D., Park, I.K., Kim, D.K.: Perceptual description for 3D object indexing and retrieval. *ISO/IEC JTC1/SC29/WG11 M8980* (2002)
2. Park, I.K., Kim, D.K., Kim, D.H., Yun, I.D., Lee, S.U.: 3D perceptual shape descriptor: Result of exploration experiments and proposal for core experiments. *ISO/IEC JTC1/SC29/WG11 M9210* (2002)
3. Park, I.K., Kim, D.K., Kim, D.H., Yun, I.D., Lee, S.U.: Perceptual 3D shape descriptor: Result of core experiment. *ISO/IEC JTC1/SC29/WG11 M9481* (2003)
4. Park, I.K., Kim, D.H., Yun, I.D., Lee, S.U.: Perceptual 3D shape descriptor: Result of core experiment. *ISO/IEC JTC1/SC29/WG11 M9809* (2003)
5. Park, I.K., Kim, D.H., Yun, I.D.: Perceptual 3D shape descriptor: Result of core experiment. *ISO/IEC JTC1/SC29/WG11 M10093* (2003)
6. Park, I.K., Kim, D.H., Yun, I.D.: Perceptual 3D shape descriptor: Result of core experiment. *ISO/IEC JTC1/SC29/WG11 M10324* (2003)
7. Yamada, A., Kim, S.K.: WD 1.0 of MPEG-7 new visual extensions. *ISO/IEC JTC1/SC29/WG11 N6367* (2004)
8. Manjunath, B.S., Salembier, P., Sikora, T.: *Introduction to MPEG-7*. John Wiley & Sons (2002)
9. Zaharia, T., Preteux, F.: Results of 3D shape core experiment. *ISO/IEC JTC1/SC29/WG11 M6315* (2000)
10. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. *ACM Transactions on Graphics* **21** (2002) 807–832
11. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., Jacobs, D.: A search engine for 3D models. *ACM Transactions on Graphics* **22** (2003) 83–105
12. Kazhdan, M., Chazelle, B., Dobkin, D., Funkhouser, T., Rusinkiewicz, S.: A reflective symmetry descriptor for 3D models. *Algorithmica* **38** (2003)
13. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: *Proceedings of Shape Modeling International 2004*, IEEE Computer Society (2004) 167–178
14. Kim, D.H., Yun, I.D., Lee, S.U.: A new shape decomposition scheme for graph-based representation. (to appear in *Pattern Recognition*)
15. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* **40** (2000) 99–121

News Video Summarization Based on Spatial and Motion Feature Analysis

Wen-Nung Lie and Chun-Ming Lai

Department of Electrical Engineering, National Chung Cheng University
160, San-Hsing, Ming-Hsiung, Chia-Yi, 621, Taiwan, ROC.
wnlie@ee.ccu.edu.tw

Abstract. In this paper, an efficient and effective summarization algorithm based on the extraction and analysis of spatial and motion features for MPEG news video is proposed. We focus on video feature analysis techniques based on the compressed domain (i.e., MVs and DCT coefficients), without the need of transformation back to the pixel domain. To give the viewers a quick and enough browse of the news content, we adopted a new strategy that the anchor audio is overlaid with the summarized news video. Hence, the detection of anchor shots and the summarization of news segment subject to a time-budget constraint constitute the two main works in this paper. In summarization of news segments, the Lagrangian multiplier approach was employed to build optimization in allocating time-lengths for all the segmented shots and getting the best perceived motion activity of the summarized video. Experiments show that our summarized news videos present an average MOS score of above 4.0 in a subjective test.

1 Introduction

Video summarization is a digestion process that gives viewers a quick, though incomplete, comprehension about the video content. Generally, this kind of digestion requires “domain-specific knowledge” to make the summarized video interesting and informative to the viewers. For examples, in baseball sport video, the behaviors after striking would be more attractive to the audience than the pitching part; for commercial movies, it is however hard to determine the synopsis unless the story is understood in advance; for news video, we may prefer to ignore the anchor’s images, but remain his/her oral introduction about each piece of news.

In literature, Ma et. al. [2] proposed a “video skimming” system based on a motion attention model. Three inductors are computed from the MV (motion vector) information by employing the “entropy” concept and then used to find out active regions in a frame. Moreover, motion activity of each active region is derived to figure out an MAI (Motion Attention Intensity) index for each frame. The finally summarized video can be constructed by picking out video segments that form local peaks of the MAI curve. Nam et. al. [3] constructed the video abstract by adopting the “non-linear sampling” method. Video data were

first segmented into shots, for which a “motion intensity index” is computed accordingly as the associated feature. Afterwards, each segmented shot is re-sampled, by changing the frame rates, according to this index value. Generally, more emphases are stressed on active segments by giving higher sampling rates.

The above two works considered motion features as the preference in video summarization process. However, this would not be true for all the viewers. A semi-automatic algorithm for video summarization was proposed by Oh et al. [4]. The viewers were first asked to pick out some “interesting” shots. Then feature analysis of these selected shots was performed to guide the summarization of the whole sequence. This strategy is to ensure that video digestion can be more conformable to individual viewer’s true preferences. Gong et al. [5] segmented the whole video sequence into several “clusters”, each of which contains “visually similar” shots. The longest shot from each cluster was then chosen and concatenated, in the time order, to form the summarized video.

In this paper, we establish an MPEG-4 news video summarization system based on the compressed-domain processing. Since information in the compressed-domain is by nature not so informative and homogeneous (no MVs in I-frames and no intra-coded DCT coefficients in P- and B-frames) than in the raw pixel domain, a series of pre-processing to convert each frame to a canonical form was developed. In this way, features can be homogeneously extracted from each frame and content analysis for each GOP (group of pictures) can be more accurate. Generally, a news program is featured of alternative concatenation between anchor shots and news segments. Several kinds of summary presentation were proposed. We choose to ignore the video part, but remain the audio part, of the anchor shot. The anchor’s oral part is then synchronously output with the digested news segments (with a similar time length). In this way, viewers can be able to understand each piece of news in the most efficient manner. This concept for news video summarization is illustrated in Fig.1. Hence, our work is reduced to anchor shot detection and summarization of news segments subject to a budget-constrained time allocation process.

It is difficult to summarize a video to fit all viewers’ preferences. To cope with this difficulty, we adopt a similar strategy as [3], that is, determining “time allocation” (or non-linear re-sampling) for each shot in the news segment. In other words, playback rates for each shot of the news segment are adaptively determined to fit the time-budget constraint. We also classify shots into “special” and “normal” events through the analysis of spatial and motion features. They will be given different weighting in time allocation to stress different emphases.

2 MPEG Video Pre-processing

The most important data in MPEG video bit streams include the MVs and the DCT coefficients. However, due to different coding principles, I-frames are lack of MVs and natures of DCT coefficients in P- and B-frames differ from those in I-frames (residual signal vs. original). This prevents us from computing informative features from MVs and DCT coefficients for a GOP. In this paper,

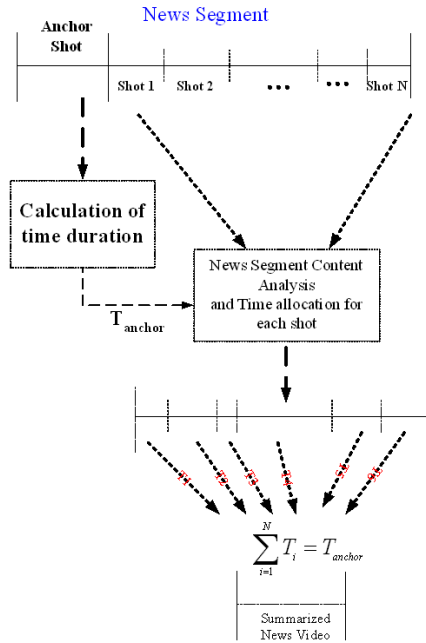


Fig. 1. Concept of proposed news video summarization process

each frame is converted into a canonical form, which is meant to have MVs as well as intra-coded DCT coefficients, regardless of the frame coding types (I-, P-, or B).

The reconstruction of intra-coded DCT coefficients for P- and B-frames can be achieved via the algorithm of motion compensation on DCT domain directly (MC-DCT) [6], without transforming DCT coefficients back to the pixel domain for motion compensation and re-performing DCT again. Considering half-pixel MVs in MPEG video, results after MC-DCT (considering only integer pixel MVs) should be refined to approximate results gained by using pixel-domain motion compensation. The algorithm proposed by Ghanbari [10] is a good choice. Based on them ([6] and [10]), accurate reconstruction of intra-coded DCT coefficients for inter-coded frames can be achieved.

As to the estimation of MVs with respect to the previous P-frame (in the preceding GOP) for I-frames, it can be achieved via interpolation by using MVs from the preceding two B-frames [7]. There are two ways: the first utilizes the “forward prediction MV” and the other utilizes “backward prediction MV”. In this step, reliability of the original MVs in B-frames is important. As is well known, MVs in MPEG videos are sometimes unreliable, since their estimation mainly aims at the least matching errors of macroblocks (MBs). Our system compares MVs between neighboring MBs and corrects those which are considered to be outliers (see [7] for details).

Actually we perform MC-DCT only for P-frames (ignoring the B-frames) based on the following three considerations: 1) feature analysis for P-frames (i.e., subsampling of frames) is enough for most of the applications, 2) eliminating the need to unifying MVs of I-, P- and B-frames will have the same reference distance (e.g., if a GOP contains 12 frames (IBBPBBPBBPBB), the reference distance is 3 frames), and 3) time-saving.

3 Anchor Shot Detection

Normally, an anchor frame can be divided into three parts: 1) anchor-man/woman, 2) news board, and 3) background. There are common features in the anchor shots: less activity, fixed composition, little camera motion, long duration, and existence of human face. In our system, the news program is assumed to start with an anchor shot. Therefore, the anchor statistics can be extracted from the first shot and used in following detection. Our anchor-shot detection algorithm is composed of two phases: shot boundary detection and shot classification.

To detect anchor shot boundaries, checks on “average color” and “histogram” differences are applied between each frame and its precedent. To do that, the DC image [1] (composed of DC terms of the DCT coefficients) is reconstructed as a reduced version (1/64 in size) of the original frame. That is, the average color and the histogram of each frame is based on its DC image only. Notice that we adopted the “perceptually weighted histogram” [8] to match two frames with translated but similar histogram shapes robustly. To have uniform color differences in matching, both computations above are based on the LUV color space, which had been suggested in MPEG-7 standard. If both the above two color checks are passed (i.e., with significant color change), the current frame is marked with “1” to identify a shot boundary candidate. According to experiments, these two checks are effective enough to segment the anchor shot precisely.

Mostly in anchor shots, active regions are located at the anchor and the news-boards. MVs could be used to find out active regions. In our system, the MPEG-7 descriptor: Motion Activity (MA, based on the magnitude of MVs) was adopted for this determination. That is, MBs having an MA larger than the threshold is recognized to belong to the active region. Inside an anchor shot, the locations of active regions would be barely changed. Hence, coordinates of its smallest enclosing rectangle are compared between two consecutive frames to determine a classification label (“1” for little change or high overlapping of enclosing rectangles and “0” otherwise).

Finally, the dominant color is defined for active regions, which is mostly found to be the color of the face or clothes of the anchor. Also, the dominant color of the active region is defined in the LUV color space. Since the first shot of the news program is assumed to be an anchor shot, we can retrieve the “dominant-color template” for subsequent shot classification and labeling.

A frame is identified as an anchor shot boundary provided that it passes both the average-color and histogram difference checks (larger than the thresholds). A segmented shot is then classified according to feature labels (for dominant color and active-region tests) of each frame contained in it. An anchor shot is identified only if enough number of positive labels exists in it.

4 Analysis and Summarization of News Segments

4.1 Shot Segmentation

Normally, the news segment can be processed as an ordinary movie and segmented into shots first. The playback speed of each shot will be adjusted in accordance with their contents. We utilize a similar method (average-color difference check) to partition each news segment. Afterwards, the segmented shots are further classified into special (flash and zoom) or normal event for different time-allocation weighting. Both the flash and zoom shots usually accompany with important events, objects, or personages and could be allocated with more time-budget.

4.2 Flash and Zoom Event Detection

Experimentally, the occurrence of flashes often results in a significant variation on the Y-component value. Hence, it is easy to classify the flash events by examining the Y-component variation between adjacent frames. The threshold is dynamically chosen to be the mean of luminance differences of the previous four frames. If any frame in the shot satisfies the check, then a flash event shot is then identified.

Zoom events can be normally detected by estimating camera motion parameters from MVs in a frame. MPEG-7 has defined several parameters concerning possible camera motions [9]. Among them, the most popular ones are: tilt, pan, roll, and zoom. Since we are merely to classify a shot as zoom shot or not, accurate estimation of the zoom parameter is unnecessary. Here, a classification, instead of estimation, problem is considered. Fig.2 demonstrates an MV field resulting from a zooming, as well as tilting and panning, camera motion. Clearly, the zoom center is not necessarily located at the center of the frame.

According to the above observations, we first train a zoom-event classifier and then use it to detect zoom events. The feature extraction and the training procedure are described below.

1. Calculate $\cos(\theta^i)$ subject to the five hypothesized zoom centers (ZC_i , $i=1\sim 5$), as shown in Fig. 3(a)(b). The angle θ^i is formed by the MV and the line connecting MB and ZC_i .

2. Calculate $E^i = \frac{1}{m \cdot n} \sum_{x=1}^m \sum_{y=1}^n w_{x,y} \cdot \cos(\theta_{(x,y)}^i)$ (x and y denote the index number of MB in the row and column directions and w is a weight to account for the reliability of the MV used to calculate θ^i) and take $i^* = \arg \max_i E^i$ as the best-fitted zoom center ZC^* . Here, we set $w_{(x,y)}$ to 1.0 for a reliable MV and

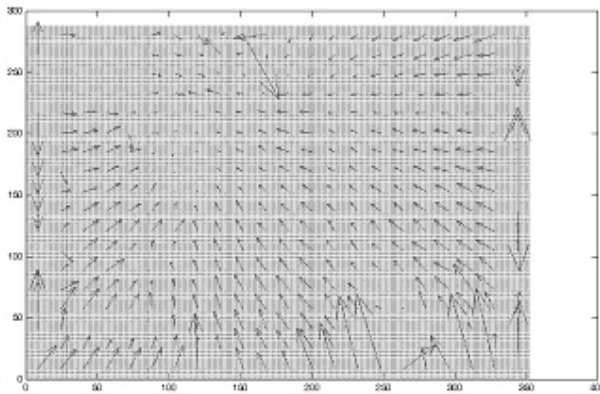


Fig. 2. Concept of proposed news video summarization process

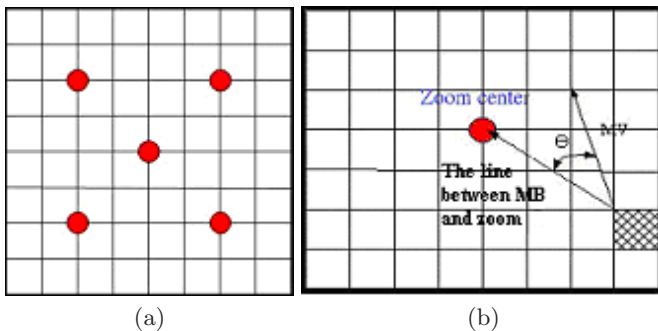


Fig. 3. (a) Five hypothesized zoom centers (b) Definition of θ

0.5 otherwise. Reliability of an MV is determined as in Section II for MV outlier correction. Notice that ZC^* is not necessarily a true one if the considered shot contains no zooming event.

3. Continue steps 1 and 2 for each frame in a shot.

4. Calculate $C_{avg} = \frac{1}{N} \sum_{j=1}^N E_j^{i^*}$ for a shot, where N denotes the number of frames in the considered shot and $E_j^{i^*}$ is computed with respect to the j -th frame's ZC^* .

5. Calculate C_{avg} for each training shot which may contain zooming event or not. Analyze C_{avg}^{zoom} 's and $C_{avg}^{non-zoom}$'s to get the optimal classification boundary zm_{thd} by minimizing the training error (number of error shots).

Normally, C_{avg} will be larger for zoom shots. In our system, we collect 60 shots (30 zoom and 30 non-zoom) and do steps 1~4 for each of them. According to experiments, $zm_{thd} = 0.497$ will cause 1 false positive and 3 false negative, i.e. the error rate is equal to 0.0667. The threshold zm_{thd} on C_{avg} can then be used

to classify the input shots. Notice that erroneous classification of shots would not mislay the summarized video, but only have an improper time allocation.

4.3 Motion-Activity-Based Time Allocation by Using Lagrangian Multiplier Optimization

To summarize the video in the news segment, our method is to allocate time lengths or determine the playback speeds for the segmented shots, subject to a total length equal to the corresponding anchor shot.

In general, viewers may pay more attention to moving objects rather than the still ones. In this paper, we would take advantage of the MVs to figure out an index of motion activity for each shot. First, the frame motion activity is defined below:

$$\text{Frame-motion-activity} = \frac{1}{m \times n} \sum_{x=1}^m \sum_{y=1}^n |mv(x, y)| \tag{1}$$

In the prior steps, we have classified each shot into “special” or “normal” group. Each group of shots will be given a different time budget (0.667 for special events and 0.333 for normal events) and optimized separately for time allocation of shots. In this paper, the Lagrangian multiplier approach is employed to cope with the issue of optimized time allocation according to shot motion-activity.

First, we define k_i (≥ 1.0) as the playback speed index for the i -th shot, N as the number of segmented shots and T_{total} as the total time budget for the considered group of event (special or normal). Time allocation is to be optimized based on the following cost function:

$$L(k_1, k_2, \dots, k_N) = f(k_1, k_2, \dots, k_N) + \lambda(g(k_1, k_2, \dots, k_N) - T_{total}) \tag{2}$$

where $f(k_1, k_2, \dots, k_N)$ and $g(k_1, k_2, \dots, k_N)$ represent the total motion activity to be maximized and the total time-length to be constrained, respectively, and λ is the multiplier. We define

$$f(k_1, k_2, \dots, k_N) = \sum_{i=1}^N (MA_i \cdot k_i) \tag{3}$$

$$g(k_1, k_2, \dots, k_N) = \sum_{i=1}^N \frac{t_{si}}{k_i} \tag{4}$$

where MA_i and t_{si} represent the original motion activity and time length of the i -th shot. MA_i is computed via

$$MA_i = \frac{1}{M} \sum_{j=1}^M \frac{1}{m \times n} \sum_{x=1}^m \sum_{y=1}^n |mv_j^i(x, y)| \tag{5}$$

where $mv_j^i(x, y)$ is the MV for the (x, y) -th MB in the j -th frame, and M is the number of frames in the i -th shot. Here, k_i can be considered as a speedup or

MA scalar. For example, the time length of the i -th shot will be reduced from t_{si} to t_{si}/k_i , while the perceived MA will be increased from MA_i to $MA_i \times k_i$. Principally, we would like to adjust the playback speeds k_i 's, according to the constraint in time T_{total} and maximization in the perceived motion activity. Shots originally with less MA would be allocated with a shorter time length, i.e. at a faster playback speed, since viewers may pay less attention to still frames.

By differentiating $L(k_1, k_2, \dots, k_N)$ with respect to k_i 's and λ , setting the results to zeros, and solving the simultaneous equations, we obtain

$$k_i^* = \frac{1}{T_{total}} \sum_{j=1}^N \sqrt{MA_j \cdot t_{sj}} \times \sqrt{\frac{t_{si}}{MA_i}} \quad (6)$$

That is, k_i^* is a function of the original time lengths t_{si} 's and the motion activities MA_i 's. Therefore, our optimized k_i^* 's reveal a balance between the time-length and perceived motion activity.

Since the structure of a GOP in MPEG standard is in the form of "IB..BPB..BPB..BPB..B", the optimized k_i^* 's should be restricted to some values so as to fit the MPEG encoding/decoding structure. In our system, we restrict k_i^* to four modes (assuming that one GOP contains 12 frames): $k_i^*=1$ (Normal play), $k_i^*=3$ (display I, P1, P2, and P3 frames only), $k_i^*=6$ (display I and P2 frames only) and $k_i^*=12$ (display I frames only).

Since the restriction on k_i^* would destroy the optimality of $L(k_1^*, k_2^*, \dots, k_N^*)$, methods other than the nearest neighbor should be explored. In our system, refinements on k_i^* 's are performed. On one hand, the finalized total time length is required to be larger and closer to T_{total} . On the other hand, the refined k_i^* values (denoted as \widehat{k}_i) should result in a least change of the originally optimized $L(k_1^*, k_2^*, \dots, k_N^*)$. Our method [7] is an iterative process that determines \widehat{k}_i 's based on a new cost function below:

$$C(\widehat{k}_1, \dots, \widehat{k}_N) = \alpha \cdot \left| \sum_{i=1}^N \frac{t_{si}}{\widehat{k}_i} - T_{total} \right| + \beta \cdot (|L(\widehat{k}_1, \dots, \widehat{k}_N) - L(k_1^*, \dots, k_N^*)|) \quad (7)$$

After the above optimization and refinement processes, a new summary video can be formed according to the \widehat{k}_i values yielded above for each piece of news shot. This can be considered as a form of video transcoding that performs news summarization. Alternatively, \widehat{k}_i 's of each shot and the shot duration (start frame and end frame) are recorded in an index file. By using a specifically designed player, the summary news video can be quickly browsed by referring to this index file and skipping the bit stream of corresponding P or B frames by simply checking the frame start codes.

5 Experimental Results

All the experiments were made with news programs from 5 TV channels. Each news program is about 5 minutes long. There are totally 22 pieces of news and

20 of them were detected, getting a recall rate of 0.91. One of the false negatives is due to a short anchor shot (we restrict an anchor shot to be above a pre-determined threshold). Since there is only 1 false alarm (resulting from a long interview whose frame composition is similar to the anchor shot), the precision rate is 0.95. We define that a correct anchor shot boundary is identified if the boundary error is within 0.5 sec (or, 15 frames). This tolerance is to account for the possibility that gradual transition often occurs in news programs. According to experiments, our algorithm was capable of correctly identifying all the anchor shots within 0.5 sec of inaccuracy.

As for the zoom event classification, we collected another 50 shots as the test set, in addition to the previously mentioned 60 training samples. Experiments found that the classification rate is 0.94 (2 false positives and 1 false negative) which is very close to the result of the training set.

We conducted some subjective tests, based on the MOS (Mean Opinion Score) measure, about the summarized news video. The MOS score ranges from 1 (lowest, poor) to 5 (highest, very good). The test procedures are as follows. First, each tested subject is trained to view some summarized videos. Then, the summarized news videos under test and their non-skimmed versions are presented subsequently. After that, the viewers are requested for scoring. By experiments, the average MOS score for 8 subjects and 12 pieces of summarized news is 4.02. Most of the unsatisfactory cases result from shorter anchor shots, which lead to over playback speed for the news segments.

On the other hand, we calculate the summarization ratio, that is, the ratio of time lengths between the summarized video and its original non-skimmed version. The values range from 0.115 to 0.336. However, the ratio depends on the news programs themselves, not on our algorithm. We also made a statistics about the accuracy in time allocation, which is defined to be the difference between the time lengths of the anchor shot and the summarized news segment. We found that the average deviation is 20 frames, i.e., 0.608 sec. This proves that our proposed algorithm for time allocation is accurate enough.

6 Conclusions

In this paper, we have proposed an efficient and effective summarization algorithm for MPEG news video based on the analysis of spatial and motion features. The MPEG video is first converted to have canonical frames, that means I-frames are augmented with MVs (with respect to the preceding P-frame) and P-frames are reconstructed with intra-coded DCT coefficients. Via this pre-processing, a set of spatial and motion features can be uniformly derived for each I- and P-frames. Notice that all the processing is efficiently conducted in the DCT domain, without transforming back to the pixel domain.

We adopted a strategy that the anchor audio is overlaid with the summarized news segment for quick browsing. Hence, two important jobs are the detection of anchor shots (so that the anchor audio can be accordingly retrieved) and summarization of the news segment subject to a time-budget constraint. The news

segment is first segmented into shots, each of which is then classified into special (flash and zoom) or normal event for different time-allocation. In the summarization of the news segments, the Lagrangian multiplier approach was employed to build optimization and make tradeoffs between the time-budget constraint and the perceived motion activity of the summarized video. Experiments show that the summarized news videos present an average MOS score of above 4.0 in a subjective test.

Future work to improve the system would be as follows.

1. Exploit techniques other than those utilizing the traditional spatial and motion information, e.g., the audio classification/recognition and face detection, in anchor shot detection to further improve the performance.
2. Expand the applications to other types of videos, e.g., sport or commercial movies.

References

1. Boon-Lock Yeo and Bede Liu: Rapid scene analysis on compressed video. *IEEE Trans. on Circuits and Systems for Video Technology*, **5** (1995) 533–544
2. Yu-Fei Ma and Hong-Jiang Zhang: A model of motion attention for video skimming. *Proc. Of IEEE Int'l Conf. on Image Processing*, (2002) 129–132
3. Jeho Nam, and A.H. Tewfik: Video Abstract of Video. *Proc. Of IEEE 3rd Workshop on Multimedia Signal Processing*, (1999) 117–122
4. JungHwan Oh, and K.A. Hua: An efficient technique for summarizing video using visual contents. *Proc. Of IEEE Int'l Conf. on Multimedia and Expo*, (2000) 1167–1170
5. Yihong Gong, and Xin Liu: Video summarization with minimal visual content redundancies. *Proc. Of IEEE Int'l Conf. on Image Processing*, (2001) 362–365
6. Shih-Fu Chang, Member, and David G. Messerschmitt: Manipulation and Compositing of MC-DCT Compressed Video. *IEEE Journal on selected areas in communications*, **13** (1995) 1–11
7. Chun-Ming Lai: MPEG-4 News Video Summarization Based on Spatial and Motion Features. Master thesis, National Chung Cheng University, Taiwan, ROC., (2003)
8. Guojun Lu and Jason Phillips: Using Perceptually Weighted Histograms for Color-based Image Retrieval. *Proc. Of Fourth Int'l Conf. on Signal Processing*, (1998) 1150–1153
9. Sylvie Jeannin, Radu Jasinschi, Alfred She, Thumpudi Naveen, Benoit Mory, and Ali Tabatabai: Motion descriptor for content-based video representation. *Signal Processing: Image Communication*, **16** (2000) 59–85
10. P.A.A. Assuncao and M. Ghanbari: A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG-2 bit streams. *IEEE Trans. On Circuits and Systems for Video Technology*, **8** (1998) 953–967

SketchIt: Basketball Video Retrieval Using Ball Motion Similarity

Sitaram Bhagavathy and Motaz El-Saban

University of California Santa Barbara, California CA 93106, USA,
{sitaram,msaban}@ece.ucsb.edu

Abstract. A prototype basketball video retrieval system is presented in this paper. Retrieval is based on the similarity of ball motion in the clip with that in the query. The system uses a query-by-sketch paradigm, where the user provides a sketch of the desired ball trajectory. The video data is pre-processed to make the ball motion invariant to camera translation. The next stage is dimensionality reduction wherein we model the ball motion as a set of parabolic trajectories. An R-tree is used to index these parabolic representations and search for similar trajectories in a low dimension parametric space. The query is processed to obtain its parametric representation, and a nearest neighbor search is performed for similar parabolas. These query results are then post-processed by assigning scores based on various similarity criteria. The system could be extended to other types of videos and moving objects. As a proof of concept, the system was tested for ball trajectories in basketball video.

1 Introduction

The wide spread use of digital media formats has made it possible to access a huge amount of content in the last few years. Nevertheless, the access methods of these different media remained the same, namely sequential access. While this approach might be acceptable if the media is not too long, or the access frequency is not too high, this scheme is highly limiting for repeated access of lengthy material. As an example, consider the task of finding small video clips that contain dunk shots from Michael Jordan in one particular NBA season. Even if the user is presented with pre-annotated content, there still remains the issue of how much time is needed for manual annotation. Several attempts have been made towards the automatic annotation of video content based on recent developments in computer vision and image understanding areas. Among those, we name here the work of Deng [Deng 97], in which video retrieval of similar content is based on a weighted similarity of color histogram, motion histogram, and Gabor texture features. [Jain 99] used a set of perceptual features to retrieve similar clips based on the extraction of a set of representative key-frames. In [Fink99], the same idea was used in retrieval, but with the features being the 2-D difference of histograms of Haar-wavelet coefficients of representative key-frames. Other approaches considered the specific problems in sports video indexing and annotation, as in [Saur 97], where basketball video is annotated by a set of labels denoting fast breaks, steals, and probable shots clips. The processing stage was fully performed in the compressed MPEG domain. [Pingali 00] developed a system for Tennis video retrieval, based on player and ball positions in the video sequence

using an automatic tracking module. All these video annotation efforts were finally standardized in [MPEG 01], which defines a set of descriptors such as color, texture, and motion. The motion descriptors in MPEG7 are different from the one we developed, because we have not indexed the trajectory points, but rather parameters derived from it. Of main importance to our work here, is the VideoQ system in [Chang 97], where they provided to the user a visual interface to sketch a query using a paint-style program. The system then retrieves video clips that contain objects with similar query attributes.

Most of the systems discussed above considered the video annotation problem rather than the indexing one. They did not pay much attention to the way the visual indexes would be stored eventually on disk. However, the storage part plays an important role in answering user queries, since a good indexing scheme means a substantial gain over the time needed to scan the whole dataset sequentially. Other researchers were focusing on the indexing problem alone. Among the most prominent schemes used for high dimensional indexing are the R-tree [Gut 84], R*-tree [Beckmann 90], k-d-tree [Bently 75], and VA-file [Weber 98]. However, one of the central issues in using indexing structures for high dimensions is the curse of dimensionality problem. So, an important step towards efficient indexing is the use of dimensionality reduction techniques to reduce the feature dimensionality. [Kobla 97] is an example system used for indexing and visualization of video sequences based on the idea of video trails. The system performed features dimensionality reduction, and then used R-tree for indexing. In [Dagtas 00], the authors presented models for motion-based video indexing for spatial and temporal invariant matches. They devised alternative computation methods for similarity measures using the fourier domain analysis techniques. An interesting work is also presented in [Perng 00] where they define similarity between time series based on a set of landmark points rather than the whole series. They also discuss how they can make their similarity measure invariant under a set of transformations such as time scaling and shifting.

Our system follows a similar approach to the work of VideoQ [Chang 97], in accepting the query as a sketch. However, the main difference between our work and their work is that they did not use a low order parametric form for the motion trajectory, which greatly simplifies the indexing part. After getting the query results from the R-tree, our system post-processes the results in order to find the most similar clips through the use of a novel score assignment technique. The rest of the paper is organized as follows. Section 2 gives the detailed description of the dimensionality reduction method used, the indexing part and the post-processing steps needed to answer queries. Section 3 gives some performance results. Finally, some conclusions and future work are presented in Section 4.

2 The System

The main objective of the system is to efficiently index basketball video clips based on their ball trajectories, for the purpose of similarity retrieval. The system should also provide an interface for the user to specify the desired query trajectory. In this section, we describe how we achieve our objective and lay out the details of the system implementation.

2.1 Data Description and Pre-processing

We obtained basketball video data from two different sources for the project. 1. A 10-minute basketball video sequence from the MPEG-7 test set. 2. A 9-minute sequence of Mr. El-Saban's basketball skills shot at the UCSB Recreation Center. From these sequences, we handpicked 87 video clips (2-4 seconds) that contained interesting ball trajectories. For each clip, using the Cognitech Video InvestigatorTM, we manually marked the basket and ball positions in each frame, to obtain two time-series of (x, y) positions. These two time-series, along with the frame numbers of the clip in the original video sequence, constitute the clip data. It should be noted that automatic tracking of the ball in a basketball game is a difficult task because of occlusions, and researchers are still trying to tackle this problem.

2.2 Dimensionality Reduction and Indexing

Consider a single frame of a video clip. Let the basket position be (x_1, y_1) and the ball position be (x_2, y_2) , as shown in Fig. 1. The top left corner is $(0, 0)$ and the y -values increase in the downward direction. We flip the frame around a vertical axis if the basket happens to lie on the right-hand side of the frame, and then fix the basket position as the origin. Now, the new ball positions are $(x_2 - x_1, y_1 - y_2)$ if the basket is on the left-hand side of the frame and $(x_1 - x_2, y_1 - y_2)$ if the basket is on the right-hand side. This approach (1) makes the ball positions invariant to camera panning, and (2) enables comparison between trajectories occurring on opposite sides of the court.

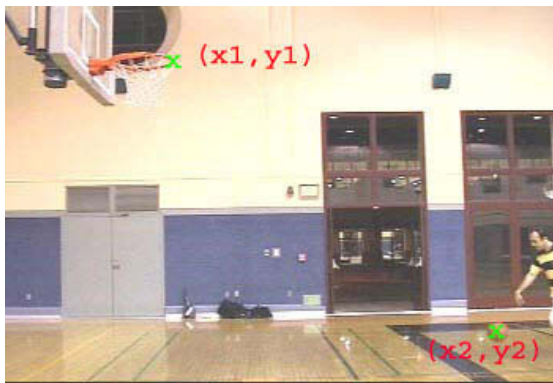


Fig. 1. A typical frame in a basket ball video clip.

Dimensionality reduction. *Principle:* The motion of rigid bodies follow parabolic trajectories in accordance with the laws of dynamics. Using the above principle, we model the ball motion in a clip as a set of parabolic trajectories. Each parabola $y = ax^2 + bx + c$ is defined by three parameters (a, b, c) . We reduce the dimensionality of the clip data as follows:

1. Pre-process the data as described earlier and obtain the new ball positions.
2. Analyze the first differences of the y -time-series to determine the time instants when the ball changed direction from downward to upward. Mark these time instants as break-points.
3. For all time instants that are not break-points yet, analyze the first differences of the x -time-series to determine if a horizontal direction change occurs. If so, mark these as break-points.
4. Consider the end-points of the time-series also as break-points. Now, divide the time-series into a set of smaller time-series, each lying between consecutive break-points.
5. Each piece is now modeled as a parabolic trajectory. This is done by fitting a parabola to the (x, y) ball position data, using a least-squares curve-fitting method.
6. The ball motion in the clip is now reduced to a set of N parabolas (a_i, b_i, c_i) , $i = 1, \dots, N$, where i is the index of the parabolas in the clip. The number of parabolas used in modelling the ball motion is based on the number of break-points found in step 2. The (a_i, b_i, c_i) and the (x, y) end-positions of each parabola are now the feature data for the clip. The feature data and the identifier (directory and file name) for each clip are stored in a feature file in order of insertion.

R-tree Indexing. Each parabola (a_i, b_i, c_i) is a point in 3-D space. These points are inserted into a R-tree index structure. The leaf nodes contain a number of parabolas, along with an object identifier for each. The index uses 20 bytes to store each leaf node entry, i.e. 12 bytes for the point, 4 bytes each for the identifier and address. The least significant 4 bits of the object identifier give the index of the parabola in the clip it came from, and the most significant 28 bits denote the byte offset of the feature data for the clip, from the beginning of the feature file. The byte offset is used to randomly access the feature data in the file.

2.3 Query Processing

Query Interface. Since the goal of the system is to support similarity retrieval, it should provide an interface for the user to specify the desired query trajectory. Our current system operates on a query-by-sketch paradigm. A query interface is provided to the user wherein the user sketches the desired query trajectory upon a static background of a basketball court. An example sketch of a query is shown in Fig. 2 The position of the basket is fixed and the sketch provides the ball positions. The (x, y) time-series are processed to obtain the query feature data, i.e. the parabolas (a_i, b_i, c_i) , $i = 1, \dots, P$, and their end-positions.

Nearest Neighbor Query. For each parabola (a_i, b_i, c_i) , $i = 1, \dots, P$, in the query, the top M matching parabolas (nearest neighbors in (a, b, c) space) are found by performing a global-order nearest neighbor search in the R-tree. The object identifier of each retrieved point leads us to the matching parabola and the clip it came from. We now have P result sets, M retrieved parabolas in each, corresponding to the P query parabolas.

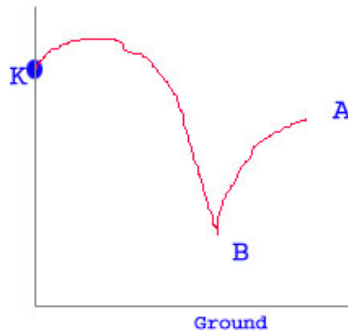


Fig. 2. A input query-by-sketch example to the system.

Post-processing. The retrievals obtained after nearest neighbor querying are processed as follows:

1. The object identifiers for each of the P result sets are compared to remove duplicate clips. This gives us the candidate set of clips.
2. For each clip in the candidate set, a score is assigned based on the satisfaction of various similarity criteria with the query sequence.
3. The clips are sorted in descending order of scores and the top K matches are returned as the similar clips.

Score computation:

Let the query be Q and a candidate clip be C . The score for this clip is computed as follows:

$$Score(Q, C) = A + B + C + D \quad (1)$$

where, A is a score assigned if the number of parabolas in the clip is the same as the number of parabolas in the query, B is a score assigned according to the number of times consecutive parabolas in the query have consecutive matches in the clip with the same order, C is a score assigned according to the similarity of the slopes and lengths of the lines of support of the query parabola and its matching parabola (the line of support for a parabola is the line joining the end points of the parabola), and D is a score assigned according to the closeness of the each query parabola with its matching parabola in the clip. The relative magnitudes of these scores are based the effect of the corresponding criterion on the perceptual similarity between the query and results. Currently, based on the quality of retrievals for various queries, we choose $A : B : C : D :: 1 : 3 : 2 : 1$.

3 Experimental Results

The original dataset consisted of 87 clips of 2-4 seconds each. From each of these clips, we generated a set of 120-125 synthetic clips by adding an offset to the x - and y -time-series and an additional uniform noise to the y -series. The final dataset contains 11074

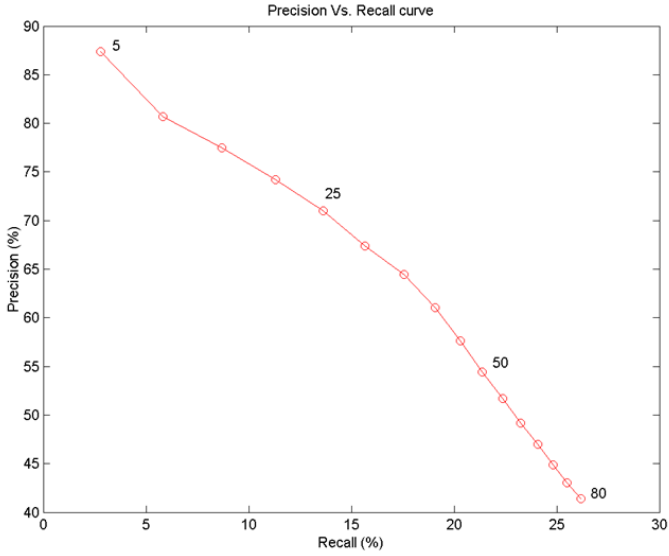
clips and includes both the original and synthetic clips. The total number of parabolas was found to be 20937. Each of these parabolas is a point in $3 - D$ space, and is inserted into an R-tree index structure along with its object identifier. Fig. 3.a shows the precision-recall curve for this dataset. The 87 original clips are used as queries and the precision and recall are averaged over all queries. The number of correct retrievals is computed by using the synthetic clips generated from each real clip, as ground-truth data for that query. The numbers on the curve stand for the number of retrievals retained after post-processing. Fig. 3.b shows the comparative query processing times for R-tree and sequential scan, at different query selectivities. The query selectivity is defined as ratio of the number of nearest neighbors (per query point) retained for post-processing, to the total number of database points. The machine used is a 667 MHz Pentium-3 PC with 512 MB RAM, running Windows 2000. The notable observations are:

1. At low selectivity, the post-processing times are negligible compared to the total query times. It increases thereafter and as the selectivity approaches 1, consumes a major chunk of the total query time.
2. At low selectivity, the total query time for an R-tree is very small compared with a sequential scan. The two times become comparable as the selectivity approaches 1.

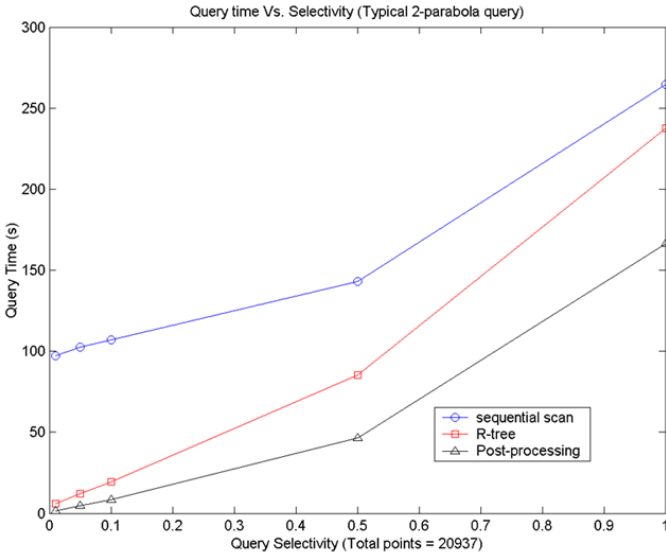
At low selectivities, the time taken for nearest neighbor querying is much larger than the post-processing time. Reducing the data dimensionality therefore results in a large improvement in the total query processing time.

4 Conclusion and Future Work

We have proposed and implemented a system that performs basketball video clip retrieval based on the similarity of ball motion with the query. The system also provides an interface for the user to sketch a desired ball trajectory by hand, on a static basketball background scene. Our pre-processing approach makes the analysis of ball trajectories invariant to camera translation, and enables the comparison of trajectories occurring at either end of the court. We then employ an innovative dimensionality reduction technique that models the ball motion in a clip as a set of parabolic trajectories. Each parabola consists of 3 parameters, and is indexed as a $3 - D$ point using an R-tree. The query is also modeled as a set of parabolas. The nearest neighbors for each query parabola comprise the candidate set for post-processing. In the post-processing, we employ a novel score assignment technique to quantify the similarity between the query and the retrieved clips. The top K matches ranked according to their scores are then returned as the retrieved clips to the user. The score assignment technique is by no means the optimal method of computing similarity between two ball trajectories. We are brainstorming for other methods that are more in accordance with perceptual similarity requirements. We are also planning to investigate whether indexing sets of two or more parabolas in a higher dimensional index structure would result in any improvement in query times and the quality of the results. An obstacle in deploying this system for general use is the requirement of manual annotation of the video clips. There is a need for effective methods for tracking the ball when there is a high probability of occlusion. We plan also to investigate other ball motion patterns that are not necessarily parabolic for a better



(a)



(b)

Fig. 3. (a) Precision-recall curve for the dataset. The numbers on the curve refer to the number of top results retained after post-processing. (b) Query processing times versus query selectivity. The circles and squares are the total query times (including post-processing) for sequential scan and R-tree, respectively. The triangles show only the post-processing times.

generalization of the system. In Finally, an important point to address in the future is the use of 3 – D trajectories. Providing the notion of depth should greatly enhance the system performance in retrieving similar motion patterns.

References

- [Deng 97] Yining Deng, and B. S. Manjunath, “Content-based Search of Video Using Color, Texture, and Motion,” Proc. IEEE International Conference on Image Processing, 1997, pp 534-537.
- [Jain 99] Anil K. Jain, Aditya Vailaya, and Xiong Wei, “Query by video clip,” *Multimedia Systems*, 7: 369-384 (1999)
- [Fink99] Xiaodong Wen, Theodore Huffmire, Helen Hu, and Adam Finkelstein, “Wavelet-based video indexing and querying,” *Multimedia Systems*, 1999, pp 350-358.
- [Saur 97] Drew Saur, Yap-Peng Tan, Sanjeev Kulkarni, and Peter Ramadge, “Automated Analysis and Annotation of Basketball Video,” Proc. SPIE, vol. 3022, 1997, pp 176-187.
- [Pingali 00] Gopal Pingali, Agata Opalach, and Yves Jean, “Ball Tracking and Virtual Replays for Innovative Tennis Broadcasts,” 2000, pp 152-156.
- [Chang 97] Shih-Fu Chang, William Chen, Horace Meng, Hari Sundaram, and Di Zhong, “A fully Automated Content Based Video Search Engine Supporting Spatio-Temporal Queries,” *ACM Multimedia*, 1997, pp 1-36.
- [Gut 84] Antonin Guttman, “R-Trees: A dynamic index structure for spatial searching”, *ACM SIGMOD*, 1984, pp 47-57.
- [Beckmann 90] Beckmann N., Kriegel H. P., Schneider R., and Seeger B., “The R*-tree: An efficient and robust access method for points and rectangles,” *Proc. ACM SIGMOD*, 1990, pp 322-331.
- [Bently 75] Bently J. L., and Friedman J. H., “Data structures for range searching,” *ACM Comput.*, pp 397-409.
- [Weber 98] Roger Weber, Hans-J. Shek, and Stephen Blott, “A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces,” *VLDB*, 1998.
- [Kobla 97] Vikrant Kobla and David Doermann, “Video Trails: Representing and Visualizing Structure in Video Sequences,” *ACM Multimedia*, 1997, pp 335-346.
- [Dagtas 00] Serhan Dagtas, Wasfi Al-Khatib, Arif Ghafoor, and R. Kashyap, “Models for motion-based video indexing and retrieval,” *IEEE Trans. on Image Processing*, Jan 2000, pp 88-101.
- [Perng 00] Chang-Shing Perng, Haixum Wang, S. Zhang, and D. Parker, “Landmarks: A new model for similarity-based pattern querying in time series databases,” *International Conference on Data Engineering*, 2000.
- [MPEG 01] B.S. Manjunath, P. Salembier, and T. Sikora, “Introduction to MPEG7: Multimedia Content Description Interface,” John Wiley and Sons Ltd., 2002.

Implanting Virtual Advertisement into Broadcast Soccer Video

Changsheng Xu, Kong Wah Wan, Son Hai Bui, and Qi Tian

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{xucs,kongwah,stushb,tian}@i2r.a-star.edu.sg

Abstract. In this paper, we propose a novel method to implant virtual advertisements into broadcast soccer video without disturbing audience's view experience. The salient objects are first detected from broadcast soccer video. These objects include static regions, central ellipse, goal mouth, and field boundary. Based on these objects, we identify suitable locations in the video for virtual advertisement implantation. The advertisement implantation can be done in real time. The proposed method allows a non-intrusive means to incorporate additional virtual content into a video presentation, facilitating an additional channel of communications to enhance greater video interactivity. It also creates a new business model for broadcast and advertising industry.

1 Introduction

The field of multimedia communications has seen tremendous growth over the past decade, leading to vast improvements that allow real-time computer-aided digital effects such as advertising image/video to be introduced into video presentations. Advertising revenues for sports sites will top USD\$6.27 billion by 2005[1]. Because sport events are often played at a stadium, coliseum or other predictable playing environment, people would expect to see predictable regions in the viewable background of a camera view that is capturing the event from a fixed position. Such regions include billboard spaces, the arena terraces, spectator stands, areas of the field, etc. Traditional advertising model in sports video is to put the advertisements in the billboards or play them during the breaks of a game. If we can create a service to dynamically implant the virtual advertisements into sports video without disturbing the view to the end viewers, it will greatly increase the advertising revenues.

The virtual advertisements are implanted in a fashion to retain a realistic view to the end viewers, so that the advertisements may be seen as appearing to be part of the scene. Once the target regions for implantation are selected, the advertisements may be selectively chosen for insertion. Audiences watching the same video broadcast in different geographical regions may then see different advertisements, advertising businesses and products that relevant to the local context.

There are several prior arts [2],[3],[4] that allow advertisements to be inserted over pre-defined fixed regions in a video broadcast. However, these approaches

need intensive labor to identify suitable target regions for virtual advertisement insertion. Once identified, these regions are fixed and no other new regions allowable. There is a need to perform virtual advertisement implantation in a realistic way in order not to disrupt the viewing experience. For instance, the virtual advertisement implanted should not obstruct the player possessing the ball during a soccer match.

There is also a conflicting requirement between the continual push for greater advertising effectiveness amongst advertisers, and the viewing pleasure of the end-viewers. Clearly, realistic virtual advertisements implanting on suitable locations are compromises enabled by current 3D graphics technology. Since there are only limited billboard spaces within the video image frames, we would expect to see advertisers pushing for more virtual spaces for advertisement implantation. The need to do this in a fashion to minimize the visual disruption to the end-viewers would be a challenge.

In this paper, we propose a novel method to implant virtual advertisements into broadcast soccer video. This provides the ultimate flexibility in assigning target regions in the video presentation for virtual content insertion and creates a new service for broadcast and advertising industry. The proposed method is fully automatic and runs in a real-time fashion, hence applicable to both video-on-demand and broadcast applications.

The rest of this paper is organized as follows. In Section 2, we present a framework of virtual advertisements implantation to sports video. In Section 3, we describe the salient objects (static region, central ellipse, goal mouth, field boundary) detection in broadcast soccer video. In Section 4, we describe the strategy to implant virtual advertisements into video. We give demonstrations and user study results in Section 5. Finally, we conclude this paper with some future work in Section 6.

2 Framework

Figure 1 illustrates the framework of virtual advertisement implantation. In order to make implanted advertisements not disrupt the viewing experience, the proper frames and regions of the video should be identified. We use some salient

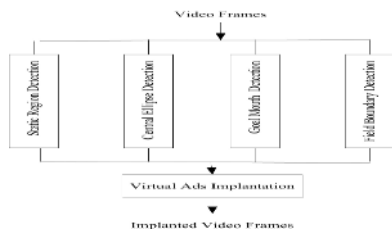


Fig. 1. Framework of virtual ads implantation

objects in sports video to identify the frames and regions for advertisements implantation. These salient objects are static regions, central ellipse, goal mouth, and field boundary. Based on the types of the detected objects, virtual advertisements are selective and implanted to the video. The salient object detection and implantation strategies will be described in detail in next 2 sections.

3 Salient Object Detection

3.1 Static Region Detection

Figure 2 illustrates the process of detecting static regions within the video frame, such as stationary TV logo, time and score bar. Each pixel in the video frame is characterized with a visual property or feature comprising of two elements: RGB intensity and directional edge strength. These are logged over a time-lag window of a pre-defined length. The pixel property change over the time-lag is recorded and its median and deviation are computed and compared against a pre-defined threshold.

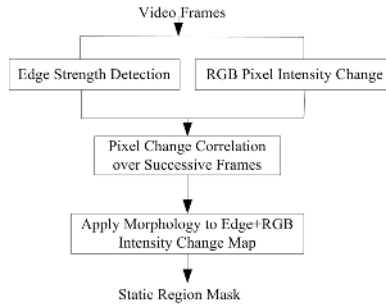


Fig. 2. Static region detection

If the change is larger than the pre-defined threshold, then the pixel is registered currently as non-static. Otherwise, it is registered as static. It is important to note that each pixel is continually analyzed for its change in order to ensure the currency of its static status registration. The reason is that these static regions may be taken off at different segments of the video, and may appear again at a later time. A different static region may also appear, at a different location. Hence, we maintain the most current set of locations where static regions are present in the video frames. Figure 3 illustrates a video frame and detected static region within the frame.

3.2 Central Ellipse Detection

We have developed a Robust Ellipse Hough Transform (REHT) [5], which is an improved ellipse Hough transform that is faster and more robust even for partial

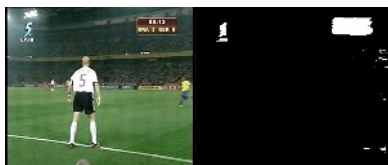


Fig. 3. Original video frame and detected static regions



Fig. 4. Original video frame and detected central ellipse

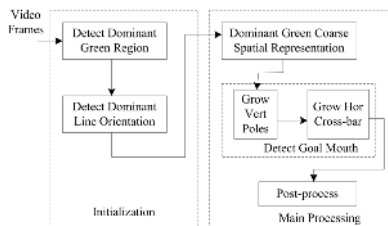


Fig. 5. System flowchart of goal mouth detection

ellipses. Our REHT is based on three main ideas: (1) a new notion of normalized measure function (NMF), which greatly simplifies the peak detection procedure, (2) a new accumulator-free computation scheme for finding the top k peaks of the NMF, without the need of a complex peak detection procedure, and (3) a generalized measure function for handling partial ellipses that make our REHT more robust. More details can be referred to [5]. Figure 4 illustrates a video frame and detected ellipse within the frame.

3.3 Goal Mouth Detection

We have developed a real-time goal mouth detection approach for broadcast soccer video [6]. This approach constraints the Hough Transform-based line-mark detection to only the dominant green regions typically seen in soccer video. The vertical goal-posts and horizontal goal-bar are isolated by color-based region growing. Figure 5 illustrates the 2 stages in goal mouth detection: (1) Initialization, which computes the key video statistics such as the dominant green and dominant goal-line orientation (we assume a dominant field color and clear line

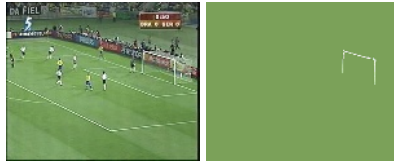


Fig. 6. Original video frame and detected goal mouth

marks), and (2) Main processing, which uses the information for goal mouth detection and segmentation. More details can be referred to [6]. Figure 6 illustrates a video frame and detected goal mouth within the frame.

3.4 Field Boundary Detection

We have developed a system [7] that tracks the camera's field-view in a soccer video in real-time. It utilizes a host of content-based visual cues that are obtained by independent threads running in parallel. A reduced resolution image is first obtained by sub-sampling the entire video frame into 32x32 non-overlapping blocks. The color distribution within each block is then examined to quantize it into either a green block or a non-green block. The green color threshold used is obtained from the parameter data set, which is itself obtained via an off-line learning process. After each block is color-quantized into green/non-green, this forms a type of *coarse color representation (CCR)* of the dominant color present in the original video frame. Since we are looking for a video frame which has the panoramic view of the field, we expect the sub-sampled coarse representation to exhibit dominantly green blocks. Therefore, connected chunks of green (non-green) blocks are computed to establish a green blob (non-green blob). A comparison of the relative size of the green blob with respect to the entire video frame size against a pre-defined threshold (also obtainable via the off-line learning process) allows the system to determine if this video frame is a field-view or not. If this frame is a field-view frame, we can detect the field boundary based on the green/non-green blob. Figure 7 illustrates a panoramic field-view frame and detected field boundary within the frame.



Fig. 7. Original video frame and detected field boundary

4 Virtual Advertisement Implantation

Based on the detected salient objects in soccer video, we come up with some strategies to identify suitable frames and locations for virtual contents implantation. Table 1 gives the statistics (percentage of duration) of the salient objects occurred in a soccer game. Since the static regions always appear in the whole video, we only select those frames in the mid-field (we use central line as a reference). The same criterion also applies to field boundary. The strategies to identify suitable locations for insertion are given as follows. For goal mouth, the virtual content is inserted above the goal bar. For central ellipse, the virtual content is scaled to ellipse shape and insert inside the central ellipse. For field boundary, virtual content is inserted outside the soccer field, like a banner. For static regions, the virtual content will cover the regions.

Table 1. Statistic of salient objects occurred in a soccer game

Central ellipse	Goal mouth	Field boundary (mid-field)	Static region (mid-field)	Total
2.4%	4.5%	8.7%	11.1%	26.7%

After the salient objects are detected and the proper locations are identified, we also provide an automatic method to select and insert the virtual content into relevant frames and regions of sports video. Depending on the geometrical property of the target regions determined by the system, a suitable form of virtual content might be implanted. For instance, if a small target region is selected, then clearly only a graphic logo can be inserted. If an entire horizontal region is deemed suitable by the system, then an animating text caption may be inserted. If a sizable target region is selected by the system, a scaled-down video insert may be used. The system also determines the exposure time duration for the virtual content. This is obtained by an off-line learning process. A similar video presentation is offered and the system learns the statistics such as shot duration, percentage of usable shots. This data is consolidated into the parameter data set to be used during actual operation.

5 Experiment

We implemented the system entirely in Win32 Visual C++ 6.0, using the open source MPEG-2 decoder (<http://libmpeg2.sourceforge.net>). Our test-set consists of 7 full games continuously recorded with the Hauppauge MPEG-1 WinTV-USB card, each coded at 1.15Mbps, CIF-352x288x25fps. We achieved a good performance on salient objects detection. For static region detection, 100% recall and precision are achieved. Above 95% recalls and precisions are achieved on central ellipse detection [5], goal mouth detection [6], and field boundary detection [7]. The virtual advertisement implantation is run in real-time on a Pentium Dell 2.2



Fig. 8. Implanted video frame samples

Ghz notebook with 1 GB RAM. Figure 8 illustrates video frames with virtual contents implanted.

To evaluate whether the implanted contents will disrupt the viewing experience or not, we ask 20 persons (all are soccer fans) to watch the video with implantation and comment whether it is acceptable or unacceptable. To our surprise, all 20 subjects give the positive comments which imply that our proposed virtual content implantation approach is promising.

6 Conclusion

We have presented a novel approach for virtual advertisement implantation to broadcast soccer video. It allows a non-intrusive means to incorporate additional virtual content into a video presentation and creates a new service for broadcast and advertising industry.

Currently the virtual advertisement implantation is based on detected salient objects in soccer video. To make the virtual advertisement implantation more efficient and effective, in the future, we will explore to use content-based video analysis technology to track the progress of the video presentation and assign a relevance-to-viewer measure to each temporal segment and each region within an individual frame in the video. The implantation will be based on such a relevance-to-viewer measure of a frame or a region. Using soccer video as examples, it would not be unreasonable to say that the viewers are focused on the immediate area around the soccer ball, and the relevance-to-viewer measure of the content depreciates as they get further concentrically away from the ball. Another example is it would be reasonable to judge that in a scene whereby the

camera view is currently focusing on the crowd, it would be of lesser relevance to the game. A further example is in a player-substitution scene. Compared with scenes where there is high global motion, player build-up, and closer to the goal-line, these examples clearly are of lesser importance to the play.

References

1. <http://www.screendigest.com>
2. S. Das, R.J. Rosser, Y. Tan, Method of Tracking Scene Motion for Live Video Insertion Systems, US Patent 5,808,695, 1998.
3. Y. Amir, et al., Method and System for Perspectively Distorting an Image and Implanting Same into a Video Stream, US Patent 5,731,846, 1998.
4. M. Tamir, et al., Method and Apparatus for Automatic Electronic Replacement of Billboards in a Video Image, US Patent 6,292,227, 2001.
5. X. Yu, H. Leong, C. Xu, and Q. Tian, "A robust Hough transform based algorithm for partial ellipse detection in broadcast soccer video", IEEE International Conference on Multimedia & Expo, Taipei, Taiwan, 27-30 Jun, 2004.
6. K. Wan, X. Yan, X. Yu, and C. Xu, "Real-time goal-mouth detection in MPEG soccer Video", ACM International Conference on Multimedia, Berkeley, USA, 2003, 311-314.
7. K. Wan, J. Lim, C. Xu, X. Yu, "Real-time camera field-view tracking in soccer video", IEEE International Conference on Acoustics, Speech, & Signal Processing, Hong Kong, China, 7-11 Apr, 2003, Vol.3, 185-188.

Automatic Video Summarization of Sports Videos Using Metadata

Yoshimasa Takahashi, Naoko Nitta, and Noboru Babaguchi

Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871 Japan
{takahashi,naoko,babaguchi}@nanase.comm.eng.osaka-u.ac.jp

Abstract. Video summarization is defined as creating a shorter video clip or a video poster which includes all but only the important scenes in an original video stream. In this paper, we propose two methods of generating a summary of arbitrary length for sports videos. One is to create a concise video clip by temporally compressing the amount of the video data. The other is to provide a video poster by spatially presenting the image keyframes which represent the whole video content. Both methods deal with the videos with metadata to summarize the video semantically. We experimentally verified the effectiveness of our method by comparing the results with man-made video summaries as well as by conducting the questionnaires to the users.

Keywords: Video Summarization, Metadata, Video Clip, Video Poster.

1 Introduction

In recent years, the development of technology has diversified service in the multimedia content, especially the video media. Thus, we need some technology to quickly search the information we want. Video summarization is one of the most promising technologies. The objective of video summarization is to create a shorter video clip or a video poster that maintains as much semantic content of the original video stream. Since semantic content of videos is difficult to automatically extract, as an alternative way, we focus on videos with their metadata, which describes the content of videos. It will greatly help the automatic generation of a video summary.

In this paper, we deal with sports videos with MPEG-7 descriptions[1,2]. MPEG-7 has been standardized to describe the semantic information of videos as metadata. It can describe hierarchical and temporal structures and the semantic content of each video segment. A sports video can be hierarchically structured according to the sports game structure. This hierarchical structure is described with MPEG-7. Semantic information such as the players and the events is attached to each video unit of this structure as well.

Let us describe related work of video summarization. Smith et al.[3] proposed a method of generating a video skim. They extracted significant information such as keywords from the audio stream, specific objects, camera motions and scene

breaks by integrating text, audio, and image analysis. The method successfully achieved the compression ratio of about 1/20 with the essential content kept. Hanjalic[4] proposed a method for extracting highlight from a sport TV broadcast. His method did not require modeling of domain-specific events. Instead, he searched for highlights at places where strong excitement is evoked in the TV viewer by the content of a video. Peker et al.[5] presented a video skimming method where the playback speed was varied based on the visual complexity for an effective fast playback. These methods only use low-level features and do not consider the semantic content, and also the time length of the summary can not be changed freely. Moreover, since the generated summary is not semantically structured, the users still have to view the whole video to search a specific scene.

As a different form of summaries, Uchihashi et al.[6] presented a method of making video posters in which the image size can be changed according to the importance measure. Chiu et al.[7] presented Stained-Glass visualization. The idea of Stained-Glass visualization is to find regions of interest in the video and to condense the keyframes into a tightly packed layout for filling the spaces between the packed regions. Although these methods improved the layout of keyframes for visualizing video summaries, they did not consider the semantic content and the number of keyframes to be displayed can not be changed dynamically.

In this paper, we propose a content-based video summarization method for sports videos using metadata given to the video media beforehand. The quality of video summary depends on whether the information which users want is included in the summary. Therefore, in this research, we consider creating the video summary which fits the length specified by the user and includes as many important video segments as possible. We also try to generate the summary in the form of the video poster which arranges image keyframes on the 2-dimensional plane as the video summary.

2 Metadata for Sports Videos

Metadata is data to describe the content, quality, condition, and other characteristics of data and includes semantic information. MPEG-7 has been recently standardized to describe the metadata for videos. In this paper, we assume the metadata, which is described with MPEG-7, is given to video media beforehand.

A sports video can be structured based on the structure of a sports game. For example, a whole baseball game is composed of several innings, an inning is composed of several at-bats, an at-bat is composed of several plays, and a play is composed of several shots. Note that a play corresponds to a pitcher throw for baseball. These hierarchical structures are described for sports videos. Additionally, for each play, five items of information, 1) the unit type, 2) the classification, 3) the players, 4) the events, and 5) the media time are described.

An example of the metadata is shown in Fig.1. In this example, the unit type is “at-bat”, the classification is “Arias”, the players are “Arias” and “Hodges”, the events are “Two-runHomeRun” and “GoAhead”, and media time includes the start time and the time length of the unit.


```

<AudioVisualSegment>
  <StructuralUnit>
    <Name>at-bat</Name>
  </StructuralUnit>
  <TextAnnotation>
    <FreeTextAnnotation>Arias</FreeTextAnnotation>
    <StructuredAnnotation>
      <Who>
        <Name>Arias</Name>
        <Name>Hodges</Name>
      </Who>
    </StructuredAnnotation>
    <KeywordAnnotation>
      <Keyword>Two-runHomeRun</Keyword>
      <Keyword>GoAhead</Keyword>
    </KeywordAnnotation>
  </TextAnnotation>
  <MediaTime>
    <MediaRelTimePoint>PODT2H37M54S20N30F</MediaRelTimePoint>
    <MediaIncrDuration>1835</MediaIncrDuration>
  </MediaTime>
  <TemporalDecomposition>
    <AudioVisualSegment>.....</AudioVisualSegment>
    <AudioVisualSegment>.....</AudioVisualSegment>
    :
  </TemporalDecomposition>
</AudioVisualSegment>

```

Fig. 1. An example of the metadata

3 Video Summarization

A strong demand for a video is to understand the content of a video in a short amount of time. We think the solution to this is the video summarization. The video summarization is defined as creating a shorter video clip or a video poster which includes all but only the important scenes in an original video stream. Therefore, each video segment should be ranked according to its semantic importance[8]. In this section, we propose a method of making a summary by ranking video segments using metadata.

3.1 Video Segment Selection

The highlights of the game should be generated based on the significance of each play. Furthermore, plays are selected so that the time length of the video summary does not exceed the time specified by the user.

3.1.1 Significance of a Play

Each video segment is given a score based on three components: the play ranks, the play occurrence time, and the number of replays.

1) Play Ranks

In this paper, we assume that a game is played between two teams, team A and team B, and that the team's goal is to get more scores than its opponent. Under this assumption, there are three states of the game situation: 'the two teams tie,' 'team A leads,' and 'team B leads.' If a play can change the current state into

a different state, we call it a State Change Play (SCP). It is evident that SCPs are candidates of the highlights.

The ranks of various plays are defined as follows:

Rank 1: SCPs.

Rank 2: score plays except SCPs.

Rank 3: plays closely related to score plays.

Rank 4: plays with score chance.

Rank 5: plays including the big play or the last play.

Rank 6: all other plays that are not in Rank 1 to 5.

Now, s_r ($0 \leq s_r \leq 1$), the rank-based significance degree of a play p_i is defined as

$$s_r(p_i) = 1 - \alpha \cdot \frac{r_i - 1}{5} \quad (1)$$

where r_i denotes the rank of the i th play p_i and α ($0 \leq \alpha \leq 1$) is the coefficient to consider how much the difference of the rank affects the significance of the play.

2) Play Occurrence Time

The score plays which are close to the end of the game largely affect the game's outcome. Thus, such plays are of great significance. We define the occurrence-time-based significance degree of a play, s_t ($0 \leq s_t \leq 1$), as

$$s_t(p_i) = 1 - \beta \cdot \frac{N - i}{N - 1} \quad (2)$$

where N is the number of all plays and β ($0 \leq \beta \leq 1$) is the coefficient to consider how much the occurrence time affects the significance of the play.

3) Number of Replays

An important play has many replays and more important plays tend to have more replays than others. So a play which has many replays is important. We define the number-of-replays-based significance degree of a play, s_n ($0 \leq s_n \leq 1$), as

$$s_n(p_i) = 1 - \gamma \cdot \frac{n_{\max} - n_i}{n_{\max}} \quad (3)$$

where n_i denotes the number of replays of the play p_i and n_{\max} is the maximum number of n_i . Also γ ($0 \leq \gamma \leq 1$) is the coefficient to consider how much the number of replays affects the significance of the play. As a consequence, significance degree of a play p_i is given by

$$s(p_i) = s_r(p_i) \cdot s_t(p_i) \cdot s_n(p_i) \quad (4)$$

Changing the parameters of α , β , and γ enables us to control the composition of the video summary. Larger α can emphasize the significance of the play rank. The other parameters behave in a similar manner.

3.1.2 Selection of Highlights

Here, the generation of the video summary based on the significance is described. For the video clip, when the time length of a video summary L is given to the system with a function $\varphi(l(p_i))$ ($0 < \varphi(l(p_i)) \leq l(p_i)$) which changes the length of each play, the problem can be formulized as follows.

$$\begin{aligned} &\text{select subset } P' = \{p_j \mid j = 1, 2, \dots, k\} \quad (1 \leq k \leq N) \\ &\text{from play set } P = \{p_1, p_2, \dots, p_N\}, \\ &\text{subject to } \sum_{p_j \in P'} s(p_j) \longrightarrow \max \\ &\qquad \qquad \sum_{p_j \in P'} \varphi(l(p_j)) \leq L \end{aligned}$$

where N denotes the total number of plays, $s(p_i)$ denotes the significance of each play, and $l(p_i)$ denotes the time length of each play. Thus, we can define this problem as the combinational optimization problem with constrained condition. For the video poster, L denotes the area of the space which displays the image keyframes, l denotes the area of each keyframe, and φ denotes the function which changes the area of each keyframe.

Basic Method: Our method selects plays in the order of significance of the plays. First, we sort out the play set P in the order of significance. Next, we pick plays in sequential order from the first play in P until the sum of the length of the selected video segments exceeds the time specified by the user.

3.2 Visualization

The form of a video summary is classified into two kinds: a video clip and a video poster. In the video clip, the time length of the original video is temporally compressed. In the video poster, image keyframes are presented on a 2-dimensional plane. We describe the details for the video clip and the video poster as follows.

3.2.1 Video Clip

The characteristic of this method is that the total length of the summary can be flexibly changed according to the time specified by the user. We propose two more methods to select the segments based on their significance.

Greedy Method: We consider the greedy algorithm. The greedy algorithm makes the selection which always seems to be the best at the point. That is, the best selection performed locally is assumed to be also the best solution globally. Our method first arranges video segments in the order of significance per unit time. Next, the play is selected in the arranged order while the sum of the length of the selected video segments does not exceed the time specified by the user.

Play-Cut Method: According to the time specified by the user, the length of the play scene is shortened dynamically. Since the length of the play is cut

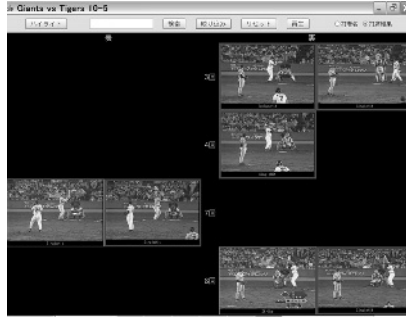


Fig. 2. The display of highlights

short, we call this method “Play-Cut Method.” The play set P is arranged in the order of significance. Next, we select plays in sequential order from the first play in P until the sum of the length of the selected video segments exceeds the time specified by the user. At this point, the following calculation is performed in order to put bounds to the length of one play.

$$\varphi(l(p_i)) = \min [l(p_i), l_{th} + \delta \cdot L'] \quad (5)$$

where l_{th} denotes the threshold of the minimum time required for the user to grasp the content of the play, and L' denotes the current remaining time after subtracting the total time of the selected plays from L , and δ is the coefficient to consider how much L' affects the length of the play. Changing the parameters of l_{th} and δ enables us to control the length of the play scenes. The longer l_{th} or δ is, the longer the length of each play becomes and consequently the number of the plays contained in the video summary decreases. When l_{th} or δ is small, although the number of the plays increases, each play segment may not fully represent the content. These parameters should be determined so that the generated summary would satisfy both requirements.

3.2.2 Video Poster

We also propose a spatial visualization system which provides image keyframes each of which represents the scene in the summary. Fig.2 shows the interface. The first frame of each scene is displayed as the keyframe. Each row represents an inning with the earliest inning being the top and the last inning being the bottom. The innings line up from top to bottom and the keyframes for each inning line up from left to right in the temporal order.

Display of Highlights: With the video poster, the users can directly specify the number of keyframes to be displayed. That is, the users can directly specify N_{new} , which denotes the new number of keyframes to be displayed. Moreover, the users can specify the ratio to the number of keyframes displayed at present by using parameter ζ ($0 \leq \zeta \leq 1$) as N_{new} . That is, N_{new} is given by

$$N_{new} = [\zeta \cdot N_f] \quad (6)$$

where N_f denotes the number of keyframes presently displayed.

Playback of each Play Scene: The user can view only the play scenes by clicking the keyframes and make the video clip suitable for his/her own preferences.

Annotations of Video Content: Since it is difficult to understand the semantic content only from the keyframes, the system also displays annotations about the players and the events under the keyframes.

4 Experimental Results

In order to verify usefulness of our system, we performed experiments and questionnaires. We prepared 5 baseball videos with the average length of 3 hours and 30 minutes. The digests which were generated by our methods were compared with the digests (of 5 games) which were actually broadcasted as the highlights on TV. We assumed that the play scenes in the digests which were broadcasted on TV are the correct answer set of the play scenes. We evaluated the results with the recall and the precision defined as follows.

$$\text{recall} = \frac{\text{number of plays included in both TV and our summary}}{\text{number of plays included in TV summary}} \times 100$$

$$\text{precision} = \frac{\text{number of plays included in both TV and our summary}}{\text{number of plays included in our summary}} \times 100$$

The value of the parameters were experimentally specified as $\alpha = 0.8$, $\beta = 0.1$, $\gamma = 0.3$, $l_{th} = 14$, $\delta = 0.02$. The comparative results between the digests generated with our methods and the digests which were broadcasted on TV are shown in Table 1. The length of the digest is set to 120 seconds for the first three methods (1. to 3. in the column “method” in Table 1) and to the same length

Table 1. Comparative Result

method	# of plays in both	# of plays on TV	# of plays in our summary	recall	precision
1.basic method	2.8	7.8	3.8	44%	80%
2.greedy method	2	7.8	8.2	29%	26%
3.play-cut method 1	5	7.8	8	66%	63%
4.play-cut method 2	5	7.8	5.8	66%	83%

Table 2. Questionnaire Result

	1	2	3	4	5	average evaluation
operativity	0%	0%	11%	78%	11%	4.0
display of highlights	0%	11%	11%	44%	33%	4.0
playback of each play scene	0%	0%	0%	22%	78%	4.8
annotations of video content	0%	0%	44%	22%	33%	3.9

(55-110 seconds) of each TV summary for the last method (4. in the column “method” in Table 1).

Then, to evaluate the effectiveness of the video poster, we gave 9 users the following questionnaires.

Q.1: Is operationality of the system good?

Q.2: Is each function convenient?

Q.3: Are there any advantages or disadvantages in the system?

The results for Q.1 and Q.2 are shown in Table 2. The users responded on a scale of 1-5 with 1 being very bad and 5 being very good. According to the responses to Q.3, the following should be considered: 1) the keyframe which better expresses the content of a play should be selected, 2) the system should be able to handle more complex queries, 3) the system should be able to retrieve only from parts of a video, etc. These are issues to be considered in the future to improve the system.

The current system serves only for baseball videos. However, the framework of the system itself is applicable to other types of sports videos with similar game structures. Moreover, since videos such as news programs have logical compositions and contain important topics and events, the system can be applied to such videos. For that purpose, the value of the parameters is required to be varied depending on the types of videos. Consequently, some systematic way to decide the value of the parameters should be contrived.

5 Conclusion

In this paper, we proposed an automatic content-based video summarization method using metadata for sports videos. We also presented two visualization systems for the video summary: video clip and video poster, formulated problems specific to each type, and proposed a method for generating the video summary of arbitrary length. As a result of experiments with baseball videos, we obtained only the significant play shots with the recall rate of 66% and the precision rate of 83% compared with the digests broadcasted on TV. As for the video poster, according to the questionnaires, the users were able to effectively access the scenes they wanted.

As a future work, semantic features should be considered to handle users' preferences. The browsing system should also be improved by extracting keyframes which best describe the semantic content of the scenes.

References

1. <http://www.chiariglione.org/mpeg/>
2. J.M.Martinez, “Overview of the MPEG-7 Standard (version 6.0),” ISO/IEC JTC1/SC29/WG11 N4509, December, 2001.
3. M.A.Smith and T.Kanade, “Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques,” Proc. IEEE CVPR'97, pp.775-781, June, 1997.

4. A.Hanjalic, "Generic Approach to Highlights Extraction from a Sport Video," Proc. IEEE ICIP 2003, Vol.1, pp.1-4, September, 2003.
5. K.A.Peker and A.Divakaran, "Adaptive Fast Playback-based Video Skimming using a Compressed-Domain Visual Complexity Measure," Proc. IEEE ICME 2004, June, 2004.
6. S.Uchihashi, J.Foote, A.Girgensohn, and J.Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries," Proc. ACM Multimedia'99, pp.383-392, October, 1999.
7. P.Chiu, A.Girgensohn, and Q.Liu, "Stained-Glass Visualization for Highly Condensed Video Summaries," Proc. IEEE ICME 2004, June, 2004.
8. N.Babaguchi, Y.Kawai, T.Ogura, and T.Kitahashi, "Personalized Abstraction of Broadcasted American Football Video by Highlight Selection," IEEE Trans. Multimedia, Vol.6, No.4, pp.575-586, August, 2004.

Classification of Frames from Broadcasted Soccer Video and Applications

Qing Tang¹, Jesse S. Jin^{1,2}, and Haiping Sun³

¹ School of Information Technologies, University of Sydney, NSW 2006, Australia
{qtang, jesse}@it.usyd.edu.au

² School of Design, Communication and I.T., University of Newcastle,
NSW 2308, Australia
jesse.jin@newcastle.edu.au

³ Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
haiping@i2r.a-star.edu.sg

Abstract. Lots of technologies have been developed for image processing. How to adapt such technologies in video analysis to achieve a balance between accuracy and computational complexity of frame analysis for video is necessary to be studied carefully. In this paper, we propose such a framework for soccer video frames. In this framework, we first pre-defined 7 classes, each of which is assigned a Semantic Descriptor (SD) to delineate its semantic meaning; then two-level analyses, local-level (called block-level) and frame-level, are done to each P frame to classify it into one of 7 categories so that the corresponding SD of a category indicates the semantic meaning of the P frame. The results, which is 80.8% in average, shows the robustness of the proposed method in accuracy; Combining with the processing speed, which is 21frames per second, shows that this framework researches the balance and is very promising.

1 Introduction

Compared with still images, videos are dynamic data with the temporal dimensions. That means a video cannot be only regarded as a sequence of still images with information in temporal dimensions ignored. While lots of techniques are developed in image retrieval, unique features of video data give rise to many new challenging issues. So how to use image analysis technologies in video analysis to increase the accuracy of frame parsing while keeping the computational complexity on a satisfactory level should be probed carefully.

M. Szummer et al. in [4] presented a method to classify images into indoor and outdoor. This is a way to combine regional analysis with global analysis for an image. A. Ekin et al. [1] divided up the screen in a 3:5:3 proportion in both directions, and positioning the main subjects on the intersection of these lines according to suggestions by A. M. Ferman et al. [2] for referee detection. This regional analysis to frames is effective with low complexity.

In this paper, we present our proposed method. We designed it for two purposes: 1) obtain satisfactory accuracy for frame analysis by using effective image

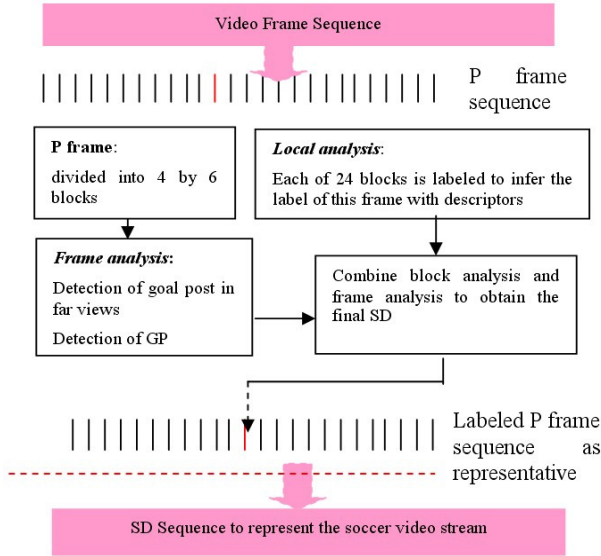


Fig. 1. System framework.

processing technologies; and 2) keep the processing speed on a nearly real-time level. The output can form the mid-level representation for soccer video for high-level event detection. The processing steps of the proposed method are shown in Figure 1.

In this method 1) We predefined 7 categories and each one with an atomic word (the SD) to indicate its semantic meaning; 2) using P frame sequence as the representative of the whole video stream; each P frame is first divided into 4 by 6 blocks and each block will be assigned one of 4 BLs to describe its semantic meaning; 3) frame-level detection of goal post; 4) according to 24 labels as well as the frame-level analysis results to infer the Semantic Descriptor to assigned to each P frame; this results in a labeled P frame sequence; 5) group the labeled P frame sequence to form the SD sequence to represent the soccer video stream.

This paper is organized as follow. In Section 2, the 7 categories for frame classification are given; in Section 3, how to assign a SD for a P frame is introduced, followed the experimental results in Section 4 and some applications based on the SD sequence in Section 5. Conclusion and future work are introduced in the last section.

2 Predefined Categories and Semantic Descriptors

Based on our observation and tests, we agree with [3] that:

1. In a soccer video, a shot may not correspond well to one semantic meaning, it could contain two or more;

Table 1. Definition of soccer video categories and corresponding Semantic Descriptors

SD	Semantic meanings	Description
CP	Close up	Close-up of a player, referee, coach, goalkeeper with no field color
AD	Audience	Far view of audience
FM	Fast movement toward a penalty box or Fight for ball control; A break happened between two penalty boxes	Far view of whole field (goal post not visible)
FP	Move inside or outside a penalty box Players are waiting for free kick or corner kick or Break	Far view of half field (goal post visible)
GP	Free kick, Corner kick, Goal, Shot or Goal Kick	Goal post in close-up view
Player(s)	Player who fouled, missed a change or take over a free kick.	Mid-range or close-up of a player
MB	Players are fighting for controlling ball.	Mid-range view (whole body visible)

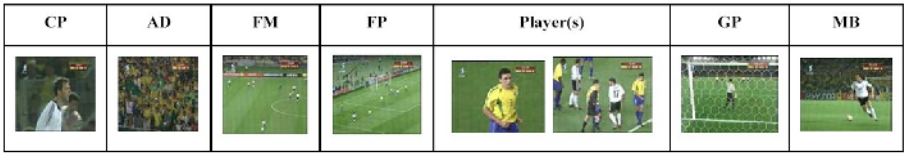


Fig. 2. Illustration of 7 SDs.

2. An shot or even the whole video stream can be decomposed as a sequence of semantically meaningful segments by using predefined semantic words (they call them visual keywords);
3. High-level video analysis can be done based on the obtained VK sequence representing the whole soccer video stream.

With these three points in mind, we proposed our method to find the balance between accuracy of frame analysis and system processing speed. Before introducing our method, the predefined 7 categories as well as their semantic meanings are given. Each P frame should be classified into one of the categories, and the semantic meaning of the category also indicates the meaning of this frame. The sample images of these 7 SDs are in Figure 2.

3 Classification Method

3.1 Block-Level Analysis

Definition of Block Label Set. In practice, we use the P frame sequence as the representative of the video stream and each P frame is divided into a 4 by



Fig. 3. Definition of Block Label Set.

6 grid, each of which is called a block. We defined 4 categories for classification of blocks, shown in Figure 3, as the Block Label Set (BLS) to delineate the semantic meanings of blocks. The analysis based on a block is called block-level analysis, and it is regional; that on the whole frame is called frame-level analysis and it is global.

Relationship between SDs and BLs. The BLS is used to label each block. Our purpose is to label each P frame with a SDs. So, the mapping relationship between the SDs and the BLS must be given. It is shown in Figure 4 and these are the ideal situations. The analysis of GP is done on frame-level, so there are 6 mapping relationships. These 6 relationships actually are 6 mapping templates.

 CP	<table border="1"> <tr><td>O</td><td>O</td><td>O</td><td>O</td><td>O</td><td>O</td></tr> <tr><td>O</td><td>O</td><td>B</td><td>B</td><td>O</td><td>O</td></tr> <tr><td>O</td><td>O</td><td>B</td><td>B</td><td>O</td><td>O</td></tr> <tr><td>O</td><td>O</td><td>B</td><td>B</td><td>O</td><td>O</td></tr> </table>	O	O	O	O	O	O	O	O	B	B	O	O	O	O	B	B	O	O	O	O	B	B	O	O	 AD	<table border="1"> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> </table>	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
O	O	O	O	O	O																																														
O	O	B	B	O	O																																														
O	O	B	B	O	O																																														
O	O	B	B	O	O																																														
A	A	A	A	A	A																																														
A	A	A	A	A	A																																														
A	A	A	A	A	A																																														
A	A	A	A	A	A																																														
 FP	<table border="1"> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> <tr><td>A</td><td>G</td><td>G</td><td>O</td><td>O</td><td>A</td></tr> <tr><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td><td>A</td></tr> <tr><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td></tr> </table>	A	A	A	A	A	A	A	G	G	O	O	A	G	G	G	G	G	A	G	G	G	G	G	G	 FM	<table border="1"> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> <tr><td>O</td><td>O</td><td>O</td><td>O</td><td>O</td><td>O</td></tr> <tr><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td></tr> <tr><td>G</td><td>G</td><td>G</td><td>B</td><td>G</td><td>G</td></tr> </table>	A	A	A	A	A	A	O	O	O	O	O	O	G	G	G	G	G	G	G	G	G	B	G	G
A	A	A	A	A	A																																														
A	G	G	O	O	A																																														
G	G	G	G	G	A																																														
G	G	G	G	G	G																																														
A	A	A	A	A	A																																														
O	O	O	O	O	O																																														
G	G	G	G	G	G																																														
G	G	G	B	G	G																																														
 Player	<table border="1"> <tr><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td><td>G</td></tr> <tr><td>G</td><td>G</td><td>B</td><td>B</td><td>G</td><td>G</td></tr> <tr><td>G</td><td>G</td><td>B</td><td>B</td><td>G</td><td>G</td></tr> <tr><td>G</td><td>G</td><td>B</td><td>B</td><td>G</td><td>G</td></tr> </table>	G	G	G	G	G	G	G	G	B	B	G	G	G	G	B	B	G	G	G	G	B	B	G	G	 MB	<table border="1"> <tr><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td><td>A</td></tr> <tr><td>A</td><td>A</td><td>A</td><td>B</td><td>B</td><td>O</td></tr> <tr><td>O</td><td>G</td><td>G</td><td>B</td><td>B</td><td>O</td></tr> <tr><td>G</td><td>G</td><td>G</td><td>B</td><td>B</td><td>G</td></tr> </table>	A	A	A	A	A	A	A	A	A	B	B	O	O	G	G	B	B	O	G	G	G	B	B	G
G	G	G	G	G	G																																														
G	G	B	B	G	G																																														
G	G	B	B	G	G																																														
G	G	B	B	G	G																																														
A	A	A	A	A	A																																														
A	A	A	B	B	O																																														
O	G	G	B	B	O																																														
G	G	G	B	B	G																																														

Fig. 4. The relationship between the Block Label Set and the Semantic Descriptors.

Feature Extraction and SD Inference. For each block of a P frame, the system extracts the following features as shown in Table 2. According to what are described in this table, the approach can automatically extract the following features. Gaussian edge detector uses $F_{edgeness}$, F_{magdir} , $Motion_{mag}$, $Motion_{dir}$, $Colors$, $FieldColor/Non - field Color$

$$F_{edgeness} = \frac{|\{P \mid Mag(p) \geq T\}|}{N} \tag{1}$$

$$F_{magdir} = (Hmag(R), Hdir(R)) \tag{2}$$

In practice, we both used a decision tree to label blocks. The decision steps are shown below.

1. In a block, if the colors are not rich and Field color is dominant, then it is labeled as “Ground”;
2. Else if its “TD” is high and “Motion” shows either stillness or movement in one direction, then it is labeled as “Audience”;
3. Else if its colors are not rich and non-field colors are dominant and motion indicates movement in different directions, it is labeled as “Body”;
4. Otherwise, it is labeled as “Other”, which means we cannot make sure its semantic meaning.

Table 2. Feature descriptions for block labels

BL	Feature Description	
Audience	Texture Density (TD)	High
	Color	Colors are rich
	Motion	Magnitude Direction
Ground	Texture Density	Low or high
	Color	Colors are not rich and field colors are dominant
	Motion	Magnitude Direction
Body	Texture Density	Low
	Color	Colors are not rich and non-field colors
	Motion	Magnitude Direction
Other	Means the type of this block can not be clearly decided	

After labeling all 24 blocks of a P frame, we calculate two features:

1. Distribution of each kind of BL on the whole frame;
2. Two thresholds were set to judge which SD should be assigned to one P frame.

3.2 Frame-Level Analysis

We hope we can detect the far view of goal post in a P frame to pick FP (far view of penalty) frame sequences from others by using line detection. The steps for detection are presented below:

1. Use Hough Transform to detect 3 white parallel lines in a P frame. The method is similar as presented in [5];

2. Use domain knowledge (two thresholds for the width and height of the white post);
3. The histograms of edge density and orientation are calculated; SVM is used as the classifier to pick up frames showing goal net from others.

We also wish to know if a P frame contains a close-up view of a goal post (GP). The method used to detect GP is also given below:

1. Use domain knowledge to detect goal post or cross bar in close-up view in each P frame. The result is A;
2. Use edge detector to detect goal net in each P frame with help of SVM. The result is B;
3. If A or B is “Yes”, then we can claim we find a goal post or cross bar in close-up view;
4. Otherwise, there is no a goal post or cross bar in close-up view detected in this frame.

Then the results from two level analyses must be combined. Suppose a P frame is labeled by “A” on block-level analysis. If the frame-level analyses labeled it as GP or FP, this P frame is finally labeled as GP or FP; otherwise, “A” is its final label.

4 Experiment Results

In order to test our method for goal post detection in close-up view, 77 video segments, including 22 of them as ground truth and 57 from other types of segments, are manually segmented and used to test the method as shown in Table 3.

Also, we manually segmented 236 clips, which comprise 102 for goal post in far view and 136 for other types of clips, to test our algorithm for goal post detection in far view in the frame-based approach. Similarly, to see if the analysis on block level can work well, we manually extracted features from 2400 blocks.

Table 3. Experimental results for GP detection

	Clips	Misclassified	Accuracy	Overall Accuracy
GP	22	3	86.4%	84.4%
Others	57	11	80.7%	

Table 4. Experimental results for FP detection

	Clips	Misclassified	Accuracy	Overall Accuracy
FP	102	6	94.1%	93.6%
Others	134	9	93.3%	

Table 5. Results for block classification

	Groundtruth	Output	Correct	Accuracy
Audience	530	472	417	88.3%
Ground	960	926	843	91.0%
Body	910	885	776	87.7%

Table 6. Experimental result

	Groundtruth	Output	Correct	Accuracy
AD	62	68	56	82.4%
FM	659	671	584	87.0%
FP	507	497	422	83.2%
MB	314	286	218	76.2%
CP	345	394	334	84.7%
GP	53	59	46	77.9%
Player	296	273	202	74.0%

Table 7. Detection of Attacking/Defending

	Groundtruth	Output	Correct	Accuracy
Attacking/Defending	47	51	44	86.3%

The experimental results are respectively presented in Table 4 and Table 5. From these results we can conclude that both of the methods can work effectively.

450 minutes with no commercial from the FIFA World Cup 2002 were used as our data set. The test results are listed in Table 6. The result, which is 80.8% in average, indicates that it research our first goal, “obtain satisfactory accuracy for frame analysis”. The processing speed is 21 frames/second, nearly real-time.

5 Related Applications

Based on the SD sequence representing the video stream, lots of high-level analysis can be done.

Except for event detection, we can define Attacking/Defending “event” for purpose of strategetic analysis or 3D reconstruction: It is a procedure in which



Fig. 5. Two ”Attacking/Defending” events.

one side gain the ball posse around mid-field and then start attacking and the other side is trying to stop the attacking. An example is given in Figure 5 to show such a procedure.

The yellow side possesses the ball and start “Attacking/Defending” from its side to white side. This procedure actually can be presented by such a SD sequence as shown in Figure 5. A coach of a soccer club can use these “Attacking/Defending” events to tell his players the problems they have in the attacking or defending; researchers can use them as well as other video streams captured by different games to construct a 3-D game.

We detect “Attacking/Defending” in the 2 of the 5 games and results are shown in Table 7.

1. FM FP FM is the basic form, and Only MB may be found in the sequence, e.g. FM MB FP FM is also expected;
2. The length of this sequence must be larger than 1 second.

6 Conclusions and Future Work

In this paper, we have presented a novel method of classifying images from soccer video stream into 7 categories, each of which has a title (SD) to describe its semantic meaning. This method contains two levels analyses; on block-level, each block of a P frame is classified in one of 4 predefined classes with the title (BL) of the class presenting its semantic meaning; then infer the SD according to the 24 BLs for this frame. Frame-level analyses are also done to form the final SD for the frame with results for block-level analyses. Based on the output SD sequence, we also discussed how to do Attacking/Defending detection as an example of high-level analysis.

In the future, we will add audio features to further improve the classification accuracy; also so, high-level event detection will be done on this SD sequence.

References

1. A. Ekin, A. M. Tekalp, R. Mehrotra: Automatic Soccer Video Analysis and Summarization. *IEEE Trans. on Image Processing*. July 2003
2. A. M. Ferman, A. M. Tekalp: A Fuzzy Framework for Unsupervised Video Content Characterization and Shot Classification. *International Journal of Electronic Imaging*. Oct. 2001
3. Haiping Sun, Joo-Hwee Lim, Qi Tian, Mohan S. Kankanhalli: Semantic Labeling of Soccer Video. *4th IEEE PCM 2003*
4. M. Szummer, R. W. Picard: Indoor-outdoor Image Classification. *IEEE International Workshop on Content-based Access of Image and Video Databases, in Conjunction with ICCV'98*
5. Kongwah Wan, Xin Yan, Xinguo Yu, Changsheng Xu: Real-Time Goal-Mouth Detection in MPEG Soccer Video. *ACM Multimedia 2003*

An Online Learning Framework for Sports Video View Classification

Jun Wu^{1,*}, XianSheng Hua², JianMin Li¹, Bo Zhang¹, and HongJiang Zhang²

¹ Department of Computer Science and Technology, Tsinghua University,
BEIJING 100084, P.R.CHINA
wujun01@mails.tsinghua.edu.cn, {lijianmin,dcszb}@mail.tsinghua.edu.cn

² Microsoft Research, Asia, 3F Sigma Center, 49 Zhichun Road,
BEIJING 100080, P.R.CHINA
{xshua,hjzhang}@microsoft.com

Abstract. Sports videos have special characteristics such as well-defined video structure, specialized sports syntax, and some canonical view types. In this paper, we proposed an online learning framework for sports video structure analysis, using baseball as an example. This framework, in which only a very small number of pre-labeled training samples are required at initial stage, employs an optimal local positive model by sufficiently exploring the local statistic characteristics of the current under-test videos. To avoid adaptive threshold selection, a set of negative models are incorporated with the local positive model during the classification procedure. Furthermore, the proposed framework is able to be applied to real time applications. Preliminary experimental results on a set of baseball videos demonstrate that the proposed system is effective and efficient.

1 Introduction

Structure analysis is an elementary step for mining the semantic information in videos. Semantic structure parsing for general videos is difficult, while for sports videos, the task is much easier due to the well-defined temporal and spatial structures in this types of videos, as well as many domain specific rules can be applied.

In [1], three types of views (global, zoom-in and closeup) of soccer games are classified according to grass ratio of the video frames, and play/break statuses are detected as basic segments using heuristic rules. Hidden Markov model and dynamic programming are applied to play/break segmentation in [2]. Gong *et al* [3] classify soccer videos into various play categories based on a prior model consisting of four major components: a soccer court, a ball, the players and the motion vectors.

In some other research efforts, sports videos are parsed by detecting canonical views, such as serve view in tennis [4][5] and pitch view in baseball [6]. In [6],

* Part of the work was performed when the author was visiting Microsoft Research Asia as a student.

a simple “play” model is defined as the basic unit, and a general sports video summarization scheme is proposed based on the definition of “play”.

Zhong *et al* [4] proposed a unified framework for scene detection and structure analysis by combining domain-specific knowledge with supervised machine learning methods. A verification step based on object and edge detection is used in order to obtain more reliable results. However, this framework has two major disadvantages. One is that good generalization capacity of the pre-trained models typically requires a sufficient amount of training data while this condition is difficult to be satisfied in most cases. The other disadvantage is that the best thresholds in classification procedure are varied for different videos. Adaptive threshold selection may help solve this issue to some extent, but the optimal thresholds are still difficult to be determined.

In this paper, a more intelligent and general online learning framework is proposed based on sufficiently exploring the local distribution properties of the video data, as well as incorporating negative models in the learning and classification procedures. This framework has the following unique features:

- 1) A dynamic local positive model is online calculated for the current under-analysis video by sufficiently exploring the local features in this video using a small amount of unlabeled samples. Furthermore, this online learning process is also able to be applied dynamically to update the local model periodically, thus making the classification results more reliable.
- 2) Negative models are sufficiently utilized to facilitate the determination of the view types, which makes the view classification results more robust and not sensitive to the thresholds.
- 3) The proposed system only requires a very small number of labeled training samples, e.g. about 10 to 40 samples.

The remainder of this paper is organized as follows. The next section is concerned with feature extraction and feature diversity analysis. In Section 3, the online learning framework is proposed in detail. Experimental results are provided in Section 4, followed by conclusion remarks and future works in the last section.

2 Feature Extraction and Diversity Analysis

Correct classification of various kinds of views in sports video is essential for further content analysis such as event detection. In this paper, pitch view in baseball is taken as an example to describe and test our system. Due to camera motion is not prominent within typical pitch view shots, as well as for the purpose of reducing computation complexity, we only use the spatial features in the key-frames of each shot. It is observed that there are special distribution characteristics for grass and sand regions in pitch views. Accordingly, in the proposed system, the feature vectors are derived from the block-wise grass and sand distribution in the key-frames.

Firstly, each key-frame is divided into $N \times N$ blocks. Then grass and sand ratios are extracted from each block respectively. It is observed that typically, the backgrounds of pitch views, such as audiences or buildings, are generally at the top of video frames, though they are varied for different stadiums or channels. Therefore, several top rows of the key-frames are ignored to make the extracted features more accurate. We define an “ignored ratio” (denoted as **IR**) as “the number of ignored rows” divided by “the total number of rows in the frame”. For example, if **IR** is set to 0.5, only $((1 - 0.5) \times N \times N)$ blocks at the bottom of the currently processing key-frame are considered, as demonstrated in Fig. 1.

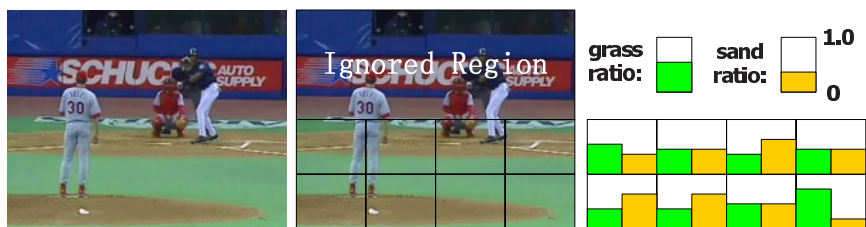


Fig. 1. Grass and sand distribution feature extraction. *Left:* an example of pitch view. *Middle:* This frame is divided into 4×4 blocks, and the two top rows are ignored (**IR** = 0.5). *Right:* grass ratio (green block) and sand ratio (yellow block) are extracted from each block, respectively.

As aforementioned, though difficult, the optimal threshold for each under-test video is necessary to be determined adaptively, in order to obtain optimal classification results. To more clearly illustrate this issue, the optimal thresholds for a view classification method based on a simple supervised learning scheme are investigated below. Five baseball videos are used in this investigation, in which 1500 shots are taken as test data. And totally there are 211 pitch view shots among them.

The major idea of the “simple learning” scheme is similar to the one in [4]. Firstly, the training samples of a specific view type are classified into a set of clusters, which results in a set of candidate view models represented by these cluster centers. Then, the key-frames of the shots in the initial segment (say, the first 10 minutes) of the under-test video are used to vote the above candidate models in order to select the best-matched model. Finally, in the classification procedure, the sample that is sufficiently close to the best-matched model is recognized as the corresponding view type.

The features used here are color histogram in HSV color space (Hereinafter referred as **hsv_36**) and grass/sand distribution (Hereinafter referred as **g_s_16**). And the performance is evaluated by $2rp/(r+p)$ since it is more discriminating than the average of p and r , where p denotes *precision* and r denotes *recall* [7]. Fig. 2 shows the two sets of curves of $2rp/(r+p)$ under different thresholds for the five videos when using **hsv_36** and **g_s_16**, respectively.

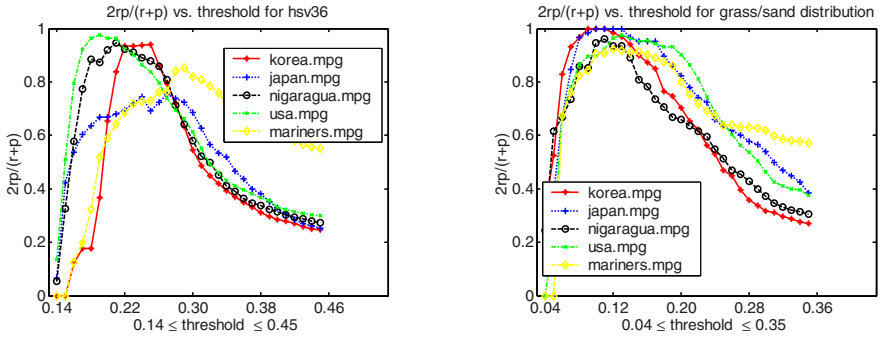


Fig. 2. Two sets of performance curves ($2rp/(r+p)$ vs. threshold) for the two types of features using the simple learning scheme. *Left:* color histogram is applied under the corresponding classification threshold ranging between 0.14 and 0.45. *Right:* grass/sand distribution feature ($N=4$ and $IR=0.5$) is employed under the classification threshold ranging between 0.04 and 0.35.

From Fig. 2, it can be seen that the grass/sand distribution feature is better than color histogram in the above learning method. And for each under-test video, the performance of view classification is basically satisfactory under the best threshold. However, the optimal thresholds for the five videos are quite different and the uniform threshold is difficult to be determined when using either feature.

As to be explained in detail in the next section, in our proposed framework, the diversity issue of the features and classification thresholds is overcome by adapting local classification models, instead of adaptively selecting optimal classification threshold.

3 An Online Learning Framework

To explore the local statistic properties of video data, we present an online learning framework for the view classification of sports videos, which dynamically learns the local statistics through the reference samples excerpted from part of current under-test video. Furthermore, to avoid adaptive threshold selection, two model sets (*candidate positive model set* and *negative model set*) are trained by a simple voting process in which only a very small number of training samples are required. In the classification procedure, both a local positive model and a set of negatives are employed. As illustrated in Fig. 3, the training process consists of five primary steps.

Step 1. Clustering “reference samples”: The key-frames of the shots from the initial part of the under-test video (unlabeled samples) are excerpted as *reference samples*. Standard K-mean clustering algorithm is employed on these samples so as to get a set of *reference clusters*, represented by their centers as $\{C_1, C_2, \dots, C_K\}$.

Step 2. Computing average inter-distance: For a specific view type, the pre-labeled training samples are clustered into another set of clusters, represented by their centers $\{S_1, S_2, \dots, S_N\}$. The average inter-distance (\bar{D}) of these clusters is defined as

$$\bar{D} = \frac{1}{N^2} \sum_{i,j=1}^N Dist(S_i, S_j), \tag{1}$$

where $Dist(S_i, S_j)$ denotes the distance between S_i and S_j . \bar{D} will be used as a distance constraint in the next step.

Step 3. Establishing “candidate positive model set”: All the pre-labeled training samples are classified into the reference clusters formed in Step 1 (by voting), among which the cluster with the largest voting number (denoted as M_0^C) is added to the candidate positive model set. And then, any other clusters that satisfy the following two conditions simultaneously are also inserted into the candidate model set.

- (a) Its voting number is larger than a certain value, e.g. one percent of the number of total training samples. Otherwise, it is regarded as noise.
- (b) The distance between the current cluster and M_0^C , $Dist(M_0^C, C_k)$, is smaller than $\alpha_1 \bar{D}$, where α_1 is a pre-defined coefficient and \bar{D} is defined in (1).

(The distance constraint $\alpha_1 \bar{D}$ is specially designed to control the size of the candidate positive model set in order to make the learned view models more compact.)

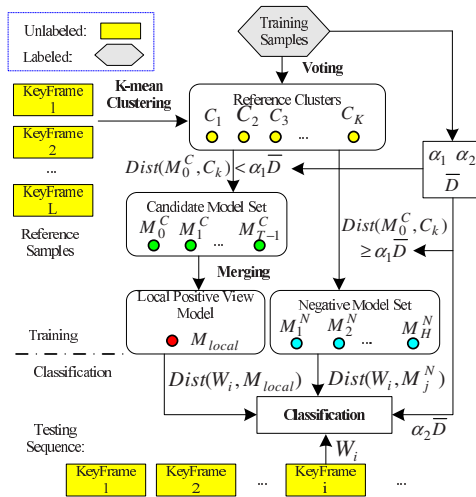


Fig. 3. Block diagram of the online learning framework. The upper part (separated by dash-dot line) shows the training procedure, and the lower part is the classification process.

Step 4. Forming “negative model set”: After the candidate positive model set is determined, all of the remaining reference clusters are considered as negative models, which are denoted as $\{M_1^N, M_2^N, \dots, M_H^N\}$.

Step 5. Calculating local model M_{local} : The local positive view model, M_{local} , is obtained by calculating the center of the merged cluster containing all the samples belonging to any of the candidate positive models.

In classification procedure, a view decision has high confidence when it is closer to the positive local model (M_{local}) than it is to the closest negative model. That is, the i -th under-test sample W_i is determined as the specific view type if

$$(i) \text{ } Dist(W_i, M_{local}) < \min_{1 \leq j \leq H} Dist(W_i, M_j^N) \text{ and}$$

(ii) $Dist(W_i, M_{local}) < \alpha_2 \bar{D}$, where α_2 is a predefined coefficient and \bar{D} has been calculated in training procedure.

The parameters in the above algorithm, α_1 and α_2 , are determined experimentally by balancing the *precision* and *recall* of view classification, as to be presented in the next section. In addition, our proposed framework is able to be applied in real time processing systems and online applications. For example, for an incoming program from the broadcasting or Internet, the first 10 minutes of video stream is used as the reference data, and after the online training procedure the bias learning framework is able to be employed to classify the follow-up video stream.

4 Experimental Results

Totally five baseball videos in MPEG-I format from different stadiums, denoted as $\{V_1, V_2, \dots, V_5\}$, are used in our experiments. In this section, firstly the selection of parameters is investigated. And then, the performance of the proposed online learning framework is evaluated, followed by some further discussions on the generalization capacity of the predefined parameters, the number of training samples, as well as the number of reference samples.

4.1 α_1 and α_2 Determination

To illustrate the generalization capacity of these parameters, the five videos are divided into two sets. One is the *validation data set*, consisting of V_1 and V_2 , from which the optimal values of the parameters, α_1 and α_2 , are determined. The other is the *test data set*, including V_3 , V_4 and V_5 , which is used to test the generalization capacity of the parameters.

As shown in Fig. 4, the highest value for average $2rp/(r+p)$ (denoted by coordinates z in the figure), 0.971, is obtained when setting $\alpha_1 = 1.3$ and $\alpha_2 = 1.0$. The performance on the test set under these optimal parameters is 0.957. It can be seen that *precision* and *recall* are not very sensitive to the parameters, especially α_1 . When α_1 is sufficiently large, the average $2rp/(r+p)$ only decreases about 2% compared with the best case. In Section 4.3, more experimental results when choosing different validation data set will be presented.

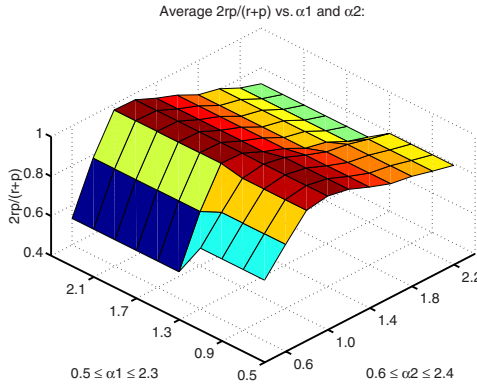


Fig. 4. Plot of the average of $2rp/(r+p)$ vs. both α_1 and α_2 upon the validation data when using grass/sand distribution as classification feature, in which α_1 ranges between 0.5 and 2.3 (coordinates y) and α_2 between 0.6 and 2.4 (coordinates x). The increasing step is 0.2.

4.2 Pitch View Classification

In this part, the performance of the proposed online learning framework is evaluated when grass/sand distribution is selected as the feature vector, and the two parameters, α_1 and α_2 , are set as the optimal values learned from the validation data set as mentioned in Section 4.1.

Aiming at simulating the circumstance of insufficient training data, while each of the five videos takes its turn acting as test data, a small quantity of pre-labeled pitch view samples in the remaining four videos will act as training data. For example, to test V_1 , the first 100 shots of V_1 are taken as the reference data and the next 300 shots as the test data. And the training data set consists of 40 pitch view samples randomly selected from the pre-labeled samples of other four videos (ten from each video).

Table 1. Pitch view classification results for two methods: the simple learning scheme and the online learning framework. The total number of the pitch view shots in the test data set is 144. For the online learning framework, $\alpha_1 = 1.3$ and $\alpha_2 = 1.0$.

Method	Feature	Training Sample No.	Error	Miss	Precision	recall
Simple Learning	hsv_36	400	13	43	0.886	0.701
Simple Learning	g_s_16	400	3	18	0.977	0.847
Online Learning	g_s_16	40	1	11	0.993	0.924

In order to compare with our proposed framework, the performance of the simple learning scheme mentioned in Section 2 is also provided in Table 1. The

classification threshold of this scheme is optimized by maximizing the classification performance on the validation data set mentioned in Section 4.1.

From the table, it can be seen that when using grass/sand distribution feature, the *precision* and *recall* of the online learning framework increases from 0.977 to 0.993 and from 0.847 to 0.924, respectively, comparing with the simple learning scheme. Furthermore, compared with color histogram, grass/sand distribution increases *precision* by 0.091 and *recall* by 0.146 for the simple learning scheme. Therefore, the grass/sand distribution feature is better than color histogram, while the online learning scheme is better than the simple learning scheme.

It should be noted that the online learning framework only uses 10% of the training samples required by the simple learning scheme, while achieves much better performance. As to be explained in Section 4.3, the number of required training samples can be further reduced without much affecting the classification performance.

4.3 Discussions for the Online Learning Framework

In order to further investigate the generalization capacity of the parameters discussed in Section 4.1, we change the sizes of the validation and test data sets simultaneously, as shown in Table 2.

Table 2. Pitch view classification performance of the online learning framework for different sizes of the validation data set and the test data set. The five videos are assigned to these two sets as shown in the following table, in which the validation data set is utilized to optimize the parameters as described in Section 4.1. The feature used here is grass/sand distribution. Note that the size of reference data is 100 shots and the total number of training samples is 40. **Opt1** means the best average $2rp/(r+p)$ on the validation data set with the optimized parameters on the validation data set listed in the table. **Opt2** denotes the highest average $2rp/(r+p)$ upon the test data set after optimizing α_1 and α_2 on the test data set.

Data Set		Training α_1, α_2			Classification Performance			
Validation	Test	α_1	α_2	Opt1	p	r	Eva	Opt2
V_1	V_2, V_3, V_4, V_5	1.3	1.0	0.958	0.988	0.928	0.957	0.978
V_1, V_2	V_3, V_4, V_5	1.3	1.0	0.971	0.993	0.924	0.957	0.974
V_1, V_2, V_3	V_4, V_5	1.3	1.2	0.970	0.990	0.922	0.955	0.973
Average Performance				0.966	0.990	0.925	0.956	0.975

The performance of the online learning framework listed in Table 2 demonstrates the robust generalization capacity for different sizes of the validation and test data set. The average classification performance on the test data sets under the parameters optimized on the validation data sets (**Eva**) is about 0.956, which is only decreased by about 0.02, compared with the average performance on the test data sets under the parameters optimized on the corresponding test data sets (**Opt2**). As a result, α_1 and α_2 have the potential to be applied to a large range of videos after pre-training on a relatively small amount of samples.

Furthermore, as aforementioned, the number of training samples can be reduced even further. According to our experiments, when it decreases from 40 to 8, the performance only descends from 0.974 to 0.972. Even only four training samples are available, average $2rp/(r+p)$ still remains at a relatively high value.

In addition, the size of the reference data currently used is 100 shots. If the application permits exploring more data, for example, the size is increased to 200 or 300 shots, the performance only improves 0.01 at most.

5 Conclusions

In this paper, we have proposed an online learning framework for sports video view classification by dynamically adapting the local classification models, in which the local statistic characteristics of the under-test videos are sufficiently explored. In this proposed framework, view classification is based on an adaptive local positive model and a set of negatives obtained by appropriately combining a small portion of unlabeled samples in the under-test video and a small amount of pre-labeled training samples, which make the results more robust and not sensitive to the parameters involved in the online learning and classification procedure. Experimental results demonstrate the effectiveness and efficiency of the proposed framework, even when only a very small quantity of training samples are available. Except for testing the proposed framework on more and wider range of videos, our future research effort will be focused on automatic feature evaluation and selection. Furthermore, an incremental learning framework, which dynamically increases the positive and negative model sets, will be explored to further improve the classification performance.

Acknowledgement. This work was supported in part by NSF Grant 60135010, 60321002.

References

1. P. Xu, *et al*, Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video, ICME, Tokyo, Aug. 22-25, 2001.
2. Lexing Xie, Shih-Fu Chang, Ajay Divakaran, Huifang Sun, Structure Analysis of Soccer Video with Hidden Markov Models, ICASSP 2002, Volume: 4, pp13-17.
3. Y.Gong, *et al*, Automatic Parsing of TV Soccer Programs, IEEE International Conference on Multimedia Computing and Systems, May, 1995, pp. 167 - 174.
4. Di Zhong, Shih-Fu Chang, Structure Analysis of Sports Video Using Domain Models, ICME, Tokyo, Aug. 22-25, 2001, pp713 -716.
5. G. Sudihir, J.C.M. Lee, A.K. Jain, Automatic classification of tennis video for high-level content-based retrieval, in Proc. IEEE International Workshop on Content-based Access of Image and Video Database, Jan, 1998, Bombay, India.
6. Baoxin Li, *et al*, Event Detection and Summarization in Sports Video, IEEE Workshop on Content based Access of Image and Video Libraries (CBAIVL), 2001
7. Raaijmakers, S.; den Hartog, J.; Baan, J.; Multimodal topic segmentation and classification of news video, ICME, Aug 26-29, 2002, pp33-36, Vol2.

A Semantic Description Scheme of Soccer Video Based on MPEG-7

Lin Liu¹, Xiuzi Ye², Min Yao³, and Sanyuan Zhang⁴

College of Computer Science / State Key Lab of CAD&CG, Hangzhou, Zhejiang
University, China
liulin@zju.edu.cn

Abstract. The birth of MPEG-7 brings an international standard for describing the content and semantic information of multimedia resource, and there are multimedia content description schemes in many areas based on MPEG-7. This paper presents a soccer video's semantic description schema based on MPEG-7 Multimedia Description Scheme. Firstly a video is segmented into shots, and then low-level features and events in shots are detected by manual or automatic methods, at last semantics of shots and the video was presented by MPEG-7. To efficiently organize the semantic information of soccer video, this paper focuses on description scheme.

Keywords: MPEG-7, Soccer Video, Video Semantic Description.

1 MPEG-7 and Multimedia Description Schemes

The MPEG-7 standard, formally named Multimedia Content Description Interface, is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group), and provides a rich set of standardized tools to describe multimedia content. Main elements of MPEG-7 standard are MPEG-7 system, Description Definition Language (DDL), MPEG-7 audio, MPEG-7 visual, Multimedia Description Schemes (MDS), MPEG-7 validation models, etc. [1][2]. Among all of these elements, MDS is a core element, which provides metadata structures for describing and annotating audio-visual (AV) content. MDS is defined using the MPEG-7 Description Definition Language (DDL) which is based on the XML Schema Language, and is instantiated as documents or streams. The definition of MDS and some description examples of MDS call be found in the file mds.xsd [3]. MDS describes AV content from several aspects, so XML files produced by MDS usually include several different type of descriptions and each of them describes one aspect of the AV content. For example, ContentEntityType description corresponds to AV content, and SummaryDescriptionType description corresponds to summary of AV content.

2 Previous Work About Soccer Video Analysis and Description

Semantic analysis of soccer video can be divided into 5 stages: the video is segmented into shots, low-level features extraction of each shot, event detection of shots, statistic analysis of the video, semantic description of the video. But all current literatures focus on the preceding 3 stages.

As to shot segmentation, methods based on color histogram or color an-glogram [4][5] prevail, and have got some satisfied result. But because of soccer video's characteristic, [7] presents a method based on color histogram and Grass-Ratio, which is more efficient in soccer video.

Low-level features have many aspects. [6][7] get shot classification according to the ratio of Grass-Color pixels to whole pixels in a frame(Grass-Ratio), [8] divided the field into 12 parts and recognized each part according to line marks and shape of the field. [9] divided a slow-motion replay into 5 parts, and detected slow-motion replay shot using a HMM model. But all these literatures neglected time information and score information.

It is difficult to automatically detect events in a shot accurately, but there is some rough method. Goal was detected according to the context of a shot, and referee was detected according to the distinguishable colored uniforms from those of two teams on the field, and penalty box was detected according to the line mark [7].

As to semantic description, automatic annotation of shot was performed in [8], and scene index was built under description of video objects [10], but they didn't conform to some specific standard. The birth of MPEG-7 presents a description scheme of multimedia content, and it makes semantic standard description of soccer video possible.

3 A Soccer Video's Semantic Description Scheme Based on MPEG-7

Video semantic information can be divided into several levels, and MPEG-7 requires the low-level features automatically detected while the high-level features allows manual input. So semantic information of soccer videos can be divided into 4 types: soccer game's context, low-level information, events information, and statistic information. Among them, soccer game's context is inputted manually, and low-level features are detected automatically, and events detection is mainly performed manually, and statistic information is calculated automatically. All information can be described with MPEG-7's MDS, and the output is XML files. The flowchart of the system is as Fig.1.

Final XML files have only 1 ContentEntityType description, which contains 2 MultimediaContent components and they are distinguished by their IDs. Among them, the component marked with ID "Full-Description" stores the soccer game's context and statistic information, and the component marked with "Segment-Description" stores low-level features and events of every shot.

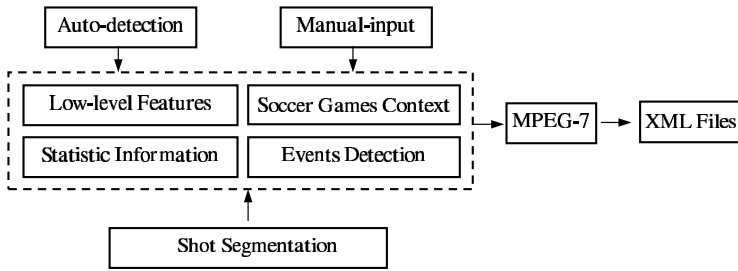


Fig. 1. The flowchart of the system

3.1 Soccer Game's Context

Context is the relative information of a soccer game, and it mainly includes following information: teams' name, time and location of the game, players' list, referees' list, uniforms color. It is difficult to detect them automatically, so they are inputted manually.

The context information is treated as text and presented by MediaInformation DS, which has two components, one is MediaIdentification DS used to locate the multi-media files, and the other is MediaProfile DS used to describe other information.

3.2 Low-Level Features

In this paper, low-level features are extracted from shots of the video.

Temporal Information. DSs describing time are based on the ISO 8601 standard which has also been adopted by the XML Schema language. Here MediaTime DS is used to describe temporal information in shots, and it records shots' start time $t1$ and duration time $t2 - t1$, in which $t2$ is end time of the shot.

Grass-Ratio. Grass-Ratio means the ratio of grass-like pixels to whole pixels of a frame, which is calculated in the key frames of every shot. Grass-Ratio can help to get the classification of a shot, the location of a shot and the current state of the game, etc.

Grass-Ratio is treated as a float value and described by SemanticState DS with ID "Grass-Ratio".

Shot Classification. A shot can be classified into several types:

Global shot: A Global shot displays the global view of the field as Fig.2. (a).

Zoom-in shot: Usually a whole human body is visible, and it is a zoom-in view of specific part of the field as Fig.2.(b).

Close-up shot: A close-up shot shows the above-waist view of one person as Fig.2.(c).

In general, an occurrence of a close-up shot indicates a break in the game. It is difficult to distinguish an out-of-field shot to a close-up shot, so an out-of-field shot is treated as a close-up shot.



Fig. 2. The classification of the shot

A shot's classification is treated as an integer value and described by Semantic-State DS with id "Shot-Class".

Playfield Zones. A soccer playfield is divided into 12 zones, 6 for each side [8]. Identification of the playfield's zones helps to detect events in the game, for example to detect corner kicks or free kicks in the game.

A playfield zone is treated by an integer value and described by SemanticState DS with id "PlayField-Zone".

Slow-Motion Replay. Replays in sports video are excellent locators for event detection. There are several types of replays, yet slow-motion replay usually occurs after shoot or serious faulty.

A slow-motion replay is treated as a Boolean value and described by SemanticState DS with id "PlayField-Zone".

Usually videos are segmented into shots, and then low-level features listed above describe every shot. In final XML files, all low-level features are stored in a Semantic DS and they are distinguished by their IDs. For example, Grass-Ratio in shot No.10 can be described as follows:

```
<Semantic id="soccer-game-seg10">
  <SemanticBase xsi:type="SemanticStateType" id="Seg0-
    Grass-Ratio">
    <Label>
      <Name> Average Grass Ratio in the
        Shot</Name>
    </Label>
    <AttributeValuePair>
      <Attribute>
        <Name>Grass-Ratio</Name>
      </Attribute>
      <FloatValue>0.583</FloatValue>
    </AttributeValuePair>
  </SemanticBase>
</Semantic>
```

3.3 Events Detection

Detection of certain events in a shot enables generation of more semantically rich description. But it is hard to automatically detect all events in a shot. In

our system, events detection is performed automatically according to low-level features and validated manually.

Shoot Detection. Shoots are detected according to the shot's context: An occurrence of shoot often brings a slow-motion replay shot and a close-up shot of players, and another feature is that a shoot shot is global shot or zoom-in shot.

Corner Kick Detection. If the start frame of a shot can detect a corner playfield zone and the shot is global shot, it indicates that maybe the shot has corner kick event.

Free Kick Detection. A free kick's character is similar to the corner kick's, but a free kick can occur everywhere in the playfield except the penalty area.

Penalty Kick Detection. Obvious the character of a penalty kick is a detection of penalty area and the playfield zone is at penalty area.

Fouls Detection. After serious fouls, the game will break and we can detect an occurrence of the referee in the shot, but slight fouls have not obvious characters and they have to be detected manually.

Events are treated as part of semantics and are described with Semantic DS, and different type events are distinguished by their IDs. For example, the shoot event of segment 101 can be described as follows:

```
<Semantic id="soccer-game-seg101">
  <SemanticBase id="Shoot-Event" xsi:type="EventType">
    <Label>
      <Name>Seg101-Shoot</Name>
    </Label>
    <AttributeValuePair>
      <Attribute>
        <Name>Shoot</Name>
      </Attribute>
      <Boolean>False</Boolean>
    </AttributeValuePair>
  </SemanticBase>
</Semantic>
```

3.4 Statistic Information

Statistic information reflects overview of the game, and includes following items:

Shot Classification Statistic Information. Includes total number of shots C_{total} , number of global shots C_{global} , number of zoom-in shots $C_{zoom-in}$, and number of slow-motion replay shots $C_{slow-mo}$.

Event Statistic Information. Includes total number of shoots E_{shoot} , total number of corner kick, free kick and penalty kick E_{free} , and total number of fouls E_{foul} .

Statistic information is treated as profile information of the game and described with MultimediaContent DS, for example, can be described as follows:

```
<MultimediaContent xsi:type="VideoType" id="Statistical-Description">
  <Video>
    <Semantic id="Soccer-Statistical-Description">
      <Label>
        <Name>Soccer-Statistical-Description</Name>
      </Label>
      <SemanticBase xsi:type="SemanticStateType">
        <Label>
          <Name>Total Shot</Name>
        </Label>
        <AttributeValuePair>
          <Attribute>
            <Name>Total Shot</Name>
          </Attribute>
          <IntegerValue>25</IntegerValue>
        </AttributeValuePair>
      </SemanticBase>
    </Semantic>
  </Video>
</MultimediaContent>
```

4 Low-Level Features Extraction

Semantic analysis of videos is based on analysis of shots, so shot segmentation is the base of semantic analysis. The first frame, the last frame and the mid frame are selected as the key frames of the shot, most low-level features are extracted from the key frames.

4.1 Shot Segmentation Based on Color Anglogram

Color anglogram is better than color histogram in shot segmentation, because color anglogram take on color statistic information and color spatial information [4][5]. Here the image is change to HSV color space, and H value and V value are chosen to make character vectors. Similarity of adjacent frames is calculated by histogram intersection method.

4.2 Playfield Zones Classification

Line Marks and shape of playfield are features used to recognize play-field zones, which are extracted from the key frames. A five elements vector is calculated from these features, which is composed by the following elements: shape descriptor F which has 6 values; the Line Mark's direction O which has 32 values; playfield size descriptor R ; playfield corner position C ; midfield line descriptor M .

A naive Byes classifier is used to classify each playfield zone Z_x , which is described in [8].

4.3 Shot Classification Based on Grass-Ratio

At first grass color on field was recognized with supervised method, then grass-ratio of every shot is calculated. According to the grass-ratio, classification of the shot can be roughly determined.

4.4 Extraction of Score Information

Score information of the game always appears at same position and with same size, color of the scoreboard is also the same. So scoreboard location can be determined with supervised method, and the score information can be extracted with the method similar to license plate recognition.

4.5 Slow-Motion Replay Detection

[9] gives us a five-stage replay model in sports games, and it can distinguish these five stages clearly. But here detection of existence of slow-motion replay is enough, so a cross-zero method [7] is adopted.

5 Experimental Results

We select 3 clips from FA Premier League 2003-2004 as experiment videos, and details of them are listed at table 1. Because these videos are long enough, we think it can represent most cases in soccer video. Accuracy of low-level features extraction is listed at table 2. All XML files produced by the system are validated through the web site described at [3].

Table 1. Details of experimental soccer video clips

Host Team	Visiting Team	Result	Clip length
Manchester United	Bolton	4:0	47:59
Arsenal	Everton	2:1	49:37
Liverpool	Chelsea	1:2	48:32

Table 2. Accuracy of low-level features extraction

Item	Total	Correct	False	Miss	Accuracy
Shot Detection	1040	920	97	23	88.4%
Shot Classification	1017	806	211	N/A	79.2%
Playfield zone classify	1017	698	319	N/A	68.6%
Slow-motion Replay Detection	93	76	20	17	81.7%

6 Conclusions and Future Work

The birth of MPEG-7 brings a standard to describe the AV semantic information. This paper present a semantic description system for soccer video and proposes a schema to describe soccer video's semantic information. The schema depends on automatic detection of low-level features and manual input of events in the video, and it gets a satisfied result and promotes the application of video library and content-based retrieval.

This paper realize a semantic description system based on MPEG-7, and the future work is to promote the accuracy of features extraction and events detection, also content-based soccer video retrieval will be studied.

References

1. Sonera Medialab:MPEG-7 White Paper. <http://www.medialab.sonera>. (2003).
2. José M. Martínez: MPEG-7 Overview (version 9) <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>.
3. <http://m7itb.nist.gov/M7Validation.html>.
4. W. Zhao, J. Wang, W. Chang: Improving Color Based Video Shot Detection. IEEE International Conference on Multimedia Computing and Systems. Volume: **2** (1999) 752–756
5. Rong Zhao, William I. Grosky: A Novel Video Shot Detection Technique Using Color Anglogram and Latent Semantic Indexing. ICDCS Workshops (2003) 550–555.
6. P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro. H. Sun: Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video, IEEE International Conference on Multimedia and Expo, Tokyo, Japan, Aug. 22-25, (2001).
7. Ahmet Ekin,A.Murat Tekalp, Rajiv Mehrotra: Automatic Soccer Video Analysis and Summarization. Image Processing, IEEE Transactions, Volume: **12**, Issue: 7 (2003) 796 – 807.
8. J.Assfalg,M.Bertini,C.Colombo,A.DelBimo, W.Nunziati: Automatic Extraction and Annotation of Soccer Video Highlights. 2003. International Conference on Image Processing. Volume: **2** (2003) 527 – 530.
9. Pan, H.; van Beek, P.; Sezan, M.I.: Detection of Slow-Motion Replay Segments in Sports Video for Highlights Generation. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). Volume: **3**,7-11 May (2001) 1649–1652.
10. Kurokawa, M.; Echigo, T.; Tomita, A.; Maeda, J.; Miyamori, H.; Iisaku, S.: Representation and retrieval of video scene by using object actions and their spatio-temporal relationships.1999 International Conference on Image Processing. ICIP 99 Proceedings. Volume: **2**, 24-28 Oct. (1999) 86–90.

Spatio-temporal Pattern Mining in Sports Video

Dong-Jun Lan¹, Yu-Fei Ma², Wei-Ying Ma², and Hong-Jiang Zhang²

¹ Dept. of Electronic Engineering, Tsinghua University, Beijing, China (100084)

landj97@mails.tsinghua.edu.cn

² Microsoft Research Asia, Sigma Center, 49 Zhichun Road, Beijing, China (100080)

{yfma, wyma, hjzhang}@microsoft.com

Abstract. Sports video is characterized with strict game rules, numerable events and well defined structures. In this paper, we proposed a generic framework for spatio-temporal pattern mining in sports video. Specifically, the periodicities in sports video are identified using unsupervised clustering and data mining method. In this way sports video analysis never needs priori domain knowledge about video genres, producers or predefined models. Therefore, such framework is easier to apply to various sports than supervised learning based approaches. In this framework, a hierarchical spatial pattern clustering routine, including scene-level clustering, field-level clustering and motion pattern clustering from top to bottom, is designed to label each subshot with coherent dominant motion. Then the temporal patterns are identified from such label sequence using data mining method. These mined probabilistic patterns are presented as basic structural elements of sports video.

1 Introduction

Sports video analysis has been addressed in many literatures. The structure analysis is aiming to facilitate structured sports video browsing. For example, in [1] basketball video is classified into wide-angle and close-up shots based on camera motion estimation. However, this simple classification cannot meet the requirement of semantic analysis. Sports video semantic analysis has primarily focused on the recognition of predefined events or actions. Semantic concept learning is presented using cues [2] or multiobjects [3][4][5] which attaches semantic concept to low-level features. In [6], a Bayesian Network approach is proposed for semantic concept network characterization. However, it is typically difficult for these approaches to extract these concepts in different conditions. As sports video is context-sensitive, many works have been done on event detection or recognition in sports video. The events in sports videos are segmented and recognized simultaneously using HMM in [7], [8]. The semantic events as well as the relationship of events are modeled by pre-trained HMM. In their method, the relationship between events is considered as grammar in sentences to improve recognition accuracy. As only motion information is taken into count, there is much room to improve, such as incorporating temporal and spatial information as a whole. Moreover, these works are confined in specific domain. Only in the

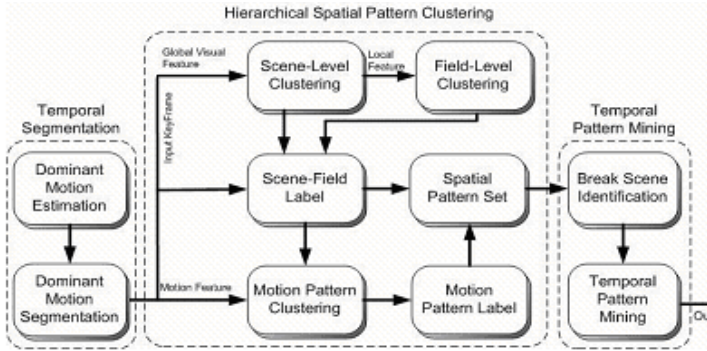


Fig. 1. System overview

condition that sports video structure is explicitly identified and model is constructed, these methods get effective. Thus generic framework for sports video analysis is desirable.

Sports video is characterized with strict game rules, numerable events and well-defined structure. In this paper, we proposed a generic framework to mine spatio-temporal patterns or periodicities in sports video, in which unsupervised clustering and data mining methods are employed. As shown in Fig. 1, for a given sports video, we segment sports video into shots and subshots according to coherent dominant motion firstly. Then a hierarchical spatial pattern clustering routine, including scene-level clustering, field-level clustering, and motion pattern clustering, are designed to mine the spatial patterns. To find the intrinsic temporal patterns in sports video, data mining methods are applied on spatial pattern sequence. Finally, those periodically appearing temporal patterns are presented as basic structural elements of sports video.

The rest of paper is organized as follows. Section 2 briefly reviews temporal segmentation. In Section 3, the hierarchical clustering routine, including scene-level clustering, field-level clustering and motion-pattern clustering, is introduced. In Section 4, a temporal pattern mining method is adopted on spatial pattern sequence. Experimental results are presented in Section 5. Section 6 concludes the paper.

2 Temporal Segmentation

In order to facilitate temporal pattern discovery, a dominant motion based segmentation method [9] is employed to segmented video sequence into shots and coherent dominant motion subshots. Integral Template Matching (*ITM*) is first applied, which is a simplified qualitative motion estimation approach. *ITM* simplified traditional motion model to 3-parameter model, which measures dominant motion in 3 independent directions, horizontal, vertical and radial. Based on the results of *ITM*, 4 quantitative and 1 qualitative curves are obtained. As shown in Fig. 2, the top three curves are the displacement curves along H , V ,

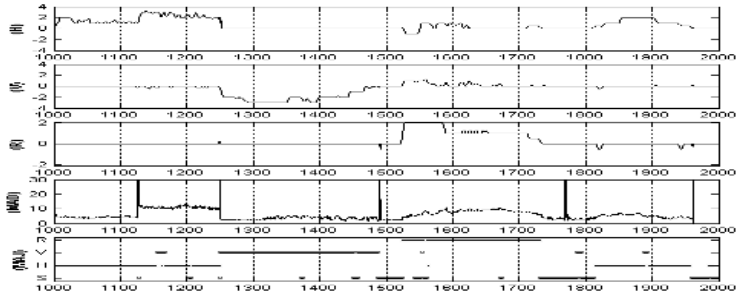


Fig. 2. Dominant motion representations

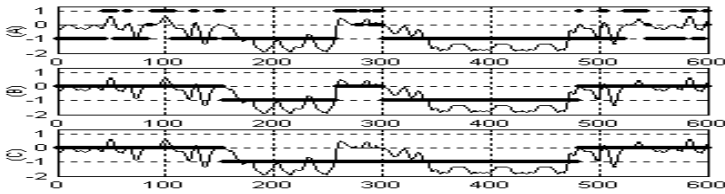


Fig. 3. Dominant motion segmentation

R respectively. The fourth curve shows minimum MAD among three directions. The last one, MAJ curve, is qualitative description of dominant motion that may be still, horizontal, vertical, or radial, determined by the minimum MAD among the three directions.

Based on the set of representation curves, we segment sports video into coherent dominant motion subshots. Morphological filtering is employed based on H , V , R curves, as shown in Fig. 3. Dominant motion subshots in sports video are classified into: pan (left/right), tilt (up/down), zoom (in/out), still, irregular motion or some motions' combination. Meanwhile, motion speed of regular motions is also given.

3 Hierarchical Spatial Pattern Clustering

In hierarchical clustering routine scene clustering is first applied using global visual feature. Then field clustering is applied to cluster fields in identical scene. Motion pattern clustering is applied using motion features and scene-field clustering results.

3.1 Scene-Level Clustering

Representative frames are first selected to represent visual content. Based on the results of temporal segmentation, we apply different strategies for different dominant motion subshots. For still subshot, the frame with the closest average

histogram is selected. For motion subshot, typically two frames are selected from the start part and end part of the subshot.

We then compute visual similarity metrics for these frames. Color and texture features are used in this stage, which represent the global visual content. Four color based features are extracted first. That is, one 256-bin *HSV* histogram in *HSV* color space and three 64-bin *Y*, *U*, *V* histograms in *YUV* color space. In addition, four correlograms are calculated, with the distance $d = 1, 3, 5, 7$ respectively. Dominant color for each frame is extracted from *HSV* histogram. To compute the similarity of these visual features, pearson's correlation coefficient is used as

$$Cor(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (1)$$

where *Cor* is from -1.0 to 1.0 . In this way we get 9-dimension similarity metrics.

The unsupervised scene-level clustering is applied on above similarity metrics. During the clustering process, the similarity measure between element and cluster, or two clusters, is defined as the average pairwise similarities of elements. Given input element E_i , the winner cluster C_I to which to assign E_i is determined through

$$C_I = \arg \left(\max_{1 \leq n \leq N} (Sim(E_i, C_n)) \right) \quad (2)$$

where N is the current cluster number. For given clustering result of k^{th} step, the clustering correlation function CCF_k is determined through the mean of average inside-cluster correlation of all current clusters,

$$CCF_k = \frac{1}{N} \sum_{1 \leq I \leq N} \left(\text{avg}_{E_i, E_j \in C_I, i \neq j} (Sim(E_i, E_j)) \right) \quad (3)$$

The clustering process can be described as following,

1. Initialize: set $N = 2$ and use first N elements E_i as initial cluster center C_i
2. For each input element E_i , assign to the winning class C_I
3. Continue step 2 until all elements have been assigned cluster label, then start following refinement, continually find the nearest C_I for each element E_i and move E_i into C_I , until no change occurring or given step arrived. Last moving information is recorded to avoid repeated moving
4. Compute clustering correlation function CCF_k of current result
5. Compute correlation increment ratio CIR_k between current clustering result and last clustering results, $CIR_k = (CSF_k - CSF_{k-1})/CSF_{k-1}$. If $CIR_k > Th$, $N = N + 1$, continue step 1-4; otherwise, stop clustering process and output clustering results.

3.2 Field-Level Clustering

Through scene clustering, different scenes in sports video are classified as shown in Fig. 4, such as close-up scene, court scene, serving scene etc. Scenes are further clustered into different fields in field-level clustering, using local visual features.

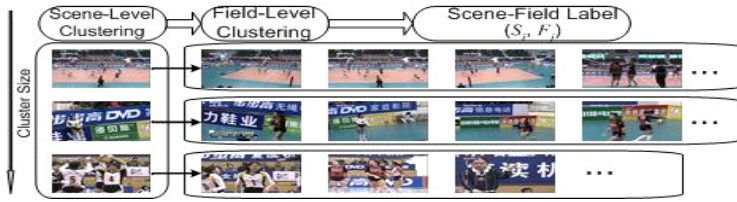


Fig. 4. Scene-level clustering and field-level clustering



Fig. 5. Motion-pattern clustering

Features extracted for field-level clustering include block histogram, edge histogram, selective block histogram, selective edge histogram and color moment. Firstly we extracted color moment, 1024-bin block histogram and 80-bin edge histogram. Selective block histogram is computed like this. The histograms of corresponding N blocks of two frames are compared. The K blocks with the largest histogram correlation are used and the similarity of the two frames is defined as the mean of the histogram correlation of the selective K blocks. In our implementation, $N = 16$ and $K = 8, 10, 12, 14$ respectively. The computation of selective edge histogram is similar to selective block histogram and $K = 13, 14, 15$. According to (1) Pearson's correlation coefficient is computed. Thus we get 10-dimension similarity metrics for clustering. The result of scene-level and field-level clustering is as Fig. 4.

3.3 Motion-Pattern Clustering

Motions are key cues for spatio-temporal pattern mining. Using scene-field clustering we assign scene-field label (S_S, F_S) , (S_E, F_E) to subshot start and end frame respectively. Motion pattern clustering is applied among those with same start and end scene label.

We extract global and object motion features to model motion patterns. For object motion, we apply block match algorithm using diamond search with block size 16×16 . Based on block-divided motion vectors mv_x , mv_y , we calculated mean and variation of motion magnitude and quantify into five levels intensity. Thus for subshot we sum 5-bin local motion mean intensity histogram and 5-bin local motion variation intensity histogram. We also calculated average motion angle and quantify into 8 types, from which we get 8-bin local motion angle histogram. For dominant motion, we get 6-bin dominant motion histogram and average motion velocity from curves in Fig. 2. Motion similarity metrics is 5-dimension and the clustering result is as Fig. 5.

4 Temporal Pattern Mining

Temporal pattern mining is applied to identify the temporal interactions and interrelations of spatial patterns in sports video. Break scene identification is first applied, because commonly play scenes are more important to constitute the semantic structure in sports video, while break scene are also used to identify start and end state set.

As shown in Fig. 4, break scenes are clustered as the results of scene-level clustering. Thus the target of break scene identification is to identify break scene clusters from all scene clusters. Break scenes are characterized with color different from field color, objects size and audio feature. Thus we firstly computed the max face region ratio based on face detection algorithm in [10]. Secondly we extract audio energy feature from audio track of sports video. In addition, from accumulative histogram of entire video, we identify dominant colors through clustering. The clustering process is similar to that of scene-level clustering. Dominant color ratio is extracted based on dominant color information. Finally, whether cluster of scene-level clustering belongs to break scene or not is decided through voting of all its cluster members.

For given spatial pattern set $P = p_1, p_2, \dots, p_N$, generated from spatial pattern clustering in section 3, temporal pattern t_i is temporal sequence constructed with P , $T_i = p_{i1}p_{i2} \dots p_{iM}$, where $Len(t_i) = M$. Temporal pattern set T_K is defined as all temporal patterns with length K , $T_k = \{t_i | t_i \in P, Len(t_i) = K\}$. For given sports video V , $V = p_{v1}p_{v2} \dots p_{vN}$, temporal pattern mining is to find temporal pattern set T_V with max probability in V . For pattern t_{vi} , $t_{vi} \in T_v$, $t_{vi} \subset V$ represent pattern t_{vi} is subsequence of V . Forward-max-match algorithm is used to search t_{vi} in V , obviously search results change with current pattern set T_V . We use $SR(t_{vi}, V)$ (Support Ratio) to measure the probability of pattern t_{vi} in V . The computation of $SR(t_{vi}, V)$ is as

$$SR(t_{vi}, V) = \frac{|\{t_{vi} | t_{vi} \subset V, t_{vi} \in T_v\}|}{Len(V)/Len(t_{vi})} \quad (4)$$

where $|\cdot|$ represent the element number of given set. For k -length pattern t_k , $T_{k+1}(t_k)$ is the pattern superset of t_k and generated from t_k , $T_{k+1}(t_k) = \{t_i | t_k \subset t_i, Len(t_i) = k + 1\}$. The data mining process is as following:

1. Construct initial state set P_s and end state set P_E
2. Initialize: set pattern length $k = 1$, for t_i in P_s , construct $T_{k+1}(t_i)$ using t_i
3. For each pattern t_j in $T_{k+1}(t_i)$, search t_j in V , compute $SR(t_j)$
4. If $SR(t_j) > Th_{SR}$, for t_j with end state in P_E , push t_j into T_V . Set $k = k + 1$, construct $T_{k+1}(t_j)$ based on t_j , search all pattern superset of t_j until pattern length $k > Th_{Len}$; if $SR(t_j) < Th_{SR}$, stop search superset of t_j , goto step 2 to search other pattern in $T_{k+1}(t_i)$
5. If candidate pattern set T_V changed, update SR of each pattern in T_V
6. For each t_{vi} in candidate pattern set T_V , if $SR(t_{vi}) > Th_{SR}$, push t_{vi} into final result set, output t_{vi} according to support ratio and pattern length.

Table 1. Temporal pattern mining results of volleyball video

Number	Temporal Pattern T_i	Total Number N_i	Support Ratio SR_i
<i>I</i>	$C - F - N - H - A$	44	0.162
<i>II</i>	$B - O - H - A$	48	0.132
<i>III</i>	$C - N - H - A$	28	0.077
<i>IV</i>	$B - G - O - H - A$	20	0.074
<i>V</i>	$F - N - H - A$	12	0.033
<i>VI</i>	$C - N - A$	12	0.022

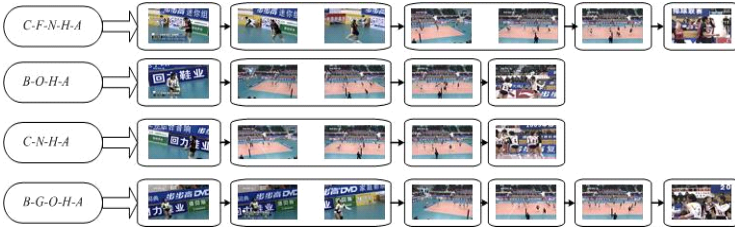


Fig. 6. Temporal pattern mining results of volleyball video

The outcome of spatio-temporal mining is a set of hierarchical spatial patterns, including scene-level, field-level and motion-level, as well as a temporal segmentation of long sports video into a set of temporal patterns in term of probability of these patterns in sports video. Based on such method, it is easy for user to grasp the semantic structure of sports video. Spatio-temporal mining also gives all related spatial or temporal patterns in sports video if users point out example. Based on the framework, we also establish domain-specific model for each sports video using HMM. The labels are used as hidden state and motion features are used as observed state. We use Torchlib to train HMM model for sports clip without manually labelling data.

5 Experiments

The proposed spatio-temporal pattern mining approach is applied to volleyball video analysis. The total duration of testing data is about 1.5 hours. These testing videos are segmented into 716 shots and 1336 subshots.

As shown in Fig. 4 and Fig. 5, based on the results of scene-level, field-level and motion pattern clustering, spatial patterns in volleyball video are grouped into 16 states. We use 16 character “*ABC...NOP*” to represent these states respectively. Thus we get a spatial pattern sequence as “*BOHABOHA BOHA...*”. Temporal mining is conducted on this sequence. The statistical results are listed in Table 1.

From Table 1, we can see that the most frequently pattern in volleyball game is C-F-N-H-A. Fig. 6 shows some examples of these mined temporal patterns.

Temporal pattern C-F-N-H-A is jump-serving of right side pattern, B-O-H-A is serving of left side pattern, C-N-H-A is serving of right side pattern, and B-G-O-H-A is jump-serving of left side pattern. Since temporal pattern mining is based on temporal statistical information, those frequently occupying patterns are identified as shown in Fig. 6. These patterns are actually the structural elements of semantics of sports video. Without priori domain knowledge about genres or producers, it proves the unsupervised method effectively finds the important segments from sports video.

6 Conclusion

In this paper, we proposed a generic framework to mine spatio-temporal patterns and periodicities in sports video using unsupervised clustering and data mining, which dose not need priori domain knowledge about sports genres or producers. Also, a hierarchical clustering routine, including scene-level clustering, field-level clustering, and motion pattern clustering (dominant and object motion), is designed to identify spatial patterns from sports video. In order to mine intrinsic temporal sports rules in sports video, a data mining method is adopted. In this manner, a set of spatio-temporal patterns are extracted from sports video without any predefined or trained model. The patterns are the primary elements for automatic sports video summarization and highlight extraction.

References

1. Y.P. Tan, D.D. Saur, S.R. Kulkarni, P.J. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation," *IEEE Trans. on Circuits and Systems for Video Technology*, vol.10, pp.133-146, 2000
2. K. Messer, W. Christmas, J. Kittler, "Automatic sports classification," *Proc. of 2002 International Conf. on Pattern Recognition*, 2002
3. M.R. Naphade, T.S. Huang, "Semantic Video Indexing Using a Probabilistic Framework," *Proc. of 2000 International Conf. on Pattern Recognition*, 2000
4. M.R. Naphade, T. Kristjansson, B.J. Frey, T.S. Huang, "Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems," *Proc. of 1998 International Conf. on Image Processing*, 1998
5. M.R. Naphade, T.S. Huang, "Semantic Video Indexing Using a Probabilistic Framework," *Proc. of 2000 International Conf. on Pattern Recognition*, 2000
6. N. Vasconcelos, A. Lippman, "A Bayesian framework for semantic content characterization," *Proc. of 2002 International Conf. on Computer Vision and Pattern Recognition*, pp. 566-571, 1998
7. Gu Xu, Y.F. Ma, H.J. Zhang, S.Q. Yang, "Motion Based Event Recognition Using HMM," *Proc. of 2002 International Conf. on Pattern Recognition*, 2002
8. Gu Xu, et al., "A HMM Based Semantic Analysis Framework For Sports Game Event Detection," *Proc. of International Conf. on Image Processing*, 2003
9. D.J. Lan, Y.F. Ma, H.J. Zhang, "A Systematic Framework of Camera Motion Analysis for Home Video," *Proc. of International Conf. on Image Processing*, 2003
10. S. Z. Li, et al., "Statistical Learning of Multi-View Face Detection," *Proc. of 2002 European Conf. on Computer Vision*, Denmark, May, 2002

Archiving Tennis Video Clips Based on Tactics Information

Jenny R. Wang¹, Nandan Prameswaran¹, Xinguo Yu²,
Changsheng Xu², and Qi Tian²

¹ School of Computer Science and Engineering,
University of New South Wales, Sydney, NSW 2052, Australia
{jennyw,paramesh}@cse.unsw.edu.au

² Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
{xinguo,xucs,tian}@i2r.a-star.edu.sg

Abstract. Video processing has found many applications in sports such as slow motion replay, pattern analysis, statistics collection, video archiving, etc. This paper presents an automatic archiving system for tennis games. It extracts ball trajectory using calibrated cameras. The trajectory is then used to classify tennis video clips into 58 tactic patterns. Semantic annotation is then attached to each clip so future query can be made easily. The annotated video clips can be used for training, sports analysis, broadcasting, etc. It can also be used for online browsing and streaming.

Keywords: Tracking, sports analysis, Bayesian networks, intelligent agents.

1 Introduction

The development of high-speed digital cameras and video processing has provoked the wide use of computers in sports. Examples include multi-camera recording and replaying [11], ball tracking [8], video analysis and summarization [3], etc. Applications have been found almost in all sports, eg, tennis [2,8,9], baseball [11,15], soccer [3], American football [7], etc. An automatic line-call system using a computer has been attempted [9,4]. Hawk-eye system has been used in Wimbledon 2003 to produce a computer-generated replay which can help the commentary team to analyze the play in eight main areas [1].

This paper attempts to archive tennis video for training purpose. It classifies tennis games into 58 winning patterns by tracking ball movement. Trajectory and landing position are used as the basic features for classification. We use an improved Bayesian networks to classify the landing position of different patterns. Intelligent agents are used to combine trajectories and landing positions as two features are in different dimensions. Semantic labels are granted after classification. The aim of the archiving is to provide a browsing tool for coaches or other personnel to retrieve tennis video clips.

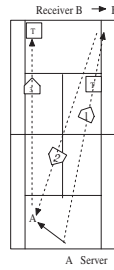


Fig. 1. Server wide to open the court pattern

2 Tennis Tactics

Tennis points are made of a sequence of shots. When a player repeat a particular sequence of shots, the game begins to take on the form of a series of patterns of play. US Tennis Association has recommended 58 winning patterns in single matches for training [13]. These patterns are based on several strategic principles that have been well tested over time. Most professional players follow these patterns and games can be classified into one of the patterns.

Fifty eight patterns are also classified into five classes:

- serve and return (18 patterns);
- groundstroke (8 patterns);
- midcourt (8 patterns);
- net play (8 patterns);
- defensive play (16 patterns).

Figure 1 shows one of the serve and return patterns on the ad court, where player A serves the ball to the far end of left court of opponent and force player B to move to the left, and then returns the ball to the far right to win the point. The list of patterns and the technical details of each pattern can be seen in USTA [13].

3 Ball Tracking and Trajectory Identification

Although Hawk-eye system has been successfully demonstrated on TV, detecting the trajectory and ball landing positions, for the accuracy sake, it is high demanding in the visible area to be within the view of at least three cameras, the cameras installed locations, the view angle of cameras, etc [4]. It constrains itself to be used in a sport studio field but widely open area. In different purpose, we detect trajectory and ball landing positions as the basic features for classification. High demanding the accuracy is unnecessary. we propose to use a wide-view camera to recover the trajectories and ball landing positions. It is easier to be installed, fewer cameras used, more economy, no obstacle to be used in widely open area, etc. The configuration and the mathematical modeling of the system are illustrated in Figure 2.

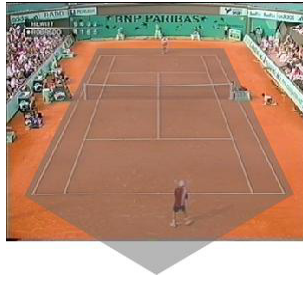


Fig. 2. Camera Settings: x-axis goes along the short side; y-axis goes along the long side.

We take the court as the global coordinate. Under the model of pinhole camera [6], we have the relation between a 3D point $\vec{u} = [u_1, u_2, u_3, 1]^T$ in the global coordinate system and its image projection $\vec{v} = [v_1, v_2, 1]^T$:

$$c\vec{v} = A[Rt]\mathbf{u} \quad (1)$$

where c is an arbitrary scale factor, extrinsic parameters $[Rt]$ is the rotation and translation which relates the global coordinate system to the camera's local coordinate system, and camera intrinsic matrix A is given by:

$$A = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

with (u_0, v_0) the coordinates of the camera principal point, α and β the scale factors in image v_1 and v_2 axes, and γ the parameter describing the skewness of the two image axes.

The camera calibration is done by manually verifying 19 corner points of the court in an image of the camera. The 19 points are five points on the net line, three points on each side service line, and four points on each side baseline. The extrinsic parameters are [14]:

$$\begin{aligned} r_1 &= \lambda A^{-1}h_1 \\ r_2 &= \lambda A^{-1}h_2 \\ r_3 &= r_1 \times r_2 \\ r_4 &= \lambda A^{-1}h_3 \end{aligned} \quad (2)$$

where $H = [h_1 h_2 h_3]^T$ is the homography which defines the relation between the 3D space and its 2D images.

After camera calibration, we obtain the value of A and $[Rt]$. By tracking the tennis ball in the video stream, we can calculate its 3D coordinate $[u_1, u_2, u_3]$. Note that u_3 is not unique because the depth information is missing in 2D projection. However, we can estimate its uniqueness through the speed of the ball. From the camera calibration, we have the speed representation in terms of

ball’s projection on the video, ie,

$$s^2 = \frac{v_x^2}{\cos^2(\alpha r_{11} + \gamma r_{21} + u_0 r_{31})} + \frac{v_y^2}{\cos^2(\alpha r_{12} + \gamma r_{22} + u_0 r_{32})} + \frac{1}{\cos^2(\alpha r_{13} + \gamma r_{23} + u_0 r_{33})}$$

The solution can be obtained using Lagrange multiplier optimization.

The trajectory in 3D has two projections $v_z^2 = p_1 v_x + p_2$ and $v_z^2 = p_3 v_y + p_4$, or in a vector form:

$$\begin{aligned} v &= \hat{v}_x \vec{p}_1 \\ v &= \hat{v}_y \vec{p}_3 \end{aligned} \tag{3}$$

where $v = v_z^2$, $\hat{v}_x = [v_x, 1]$, $\vec{p}_1 = [p_1, p_2]^T$, $\hat{v}_y = [v_y, 1]$ and $\vec{p}_3 = [p_3, p_4]^T$. From multiple sample points on the trajectory, we can obtain $\vec{p}_1 = (\hat{v}_x^T \hat{v}_x)^{-1} \hat{v}_x^T v$, and $\vec{p}_3 = (\hat{v}_y^T \hat{v}_y)^{-1} \hat{v}_y^T v$ [5]. The 3D trajectory can be obtained from (1) and (3).

4 Pattern Classification

To summarize ball’s landing positions in tennis tactics, we have 13 clusters in a half court for a deuce court player, as shown in Figure 3a. It is difficult to calculate the precise landing location of the ball to estimate the cluster it belongs to. Making more efforts to improve the accuracy would not help as the error could come from the player. We simplify clusters into five major clusters, as shown in Figure 3b. The shadow clusters are symmetrical ones for the ad court. It is much easy to distinguish between the clusters. For further clustering we use an improved Bayesian network and intelligent agents to resolve the ambiguity, which will be elaborated in the next section.

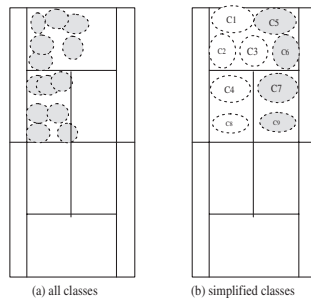


Fig. 3. Ball’s landing positions

By the same token, we summarize all tactic trajectories for a deuce court serving game, as shown in Figure 4a. It is clearly difficult for a computer to

identify and distinguish them. We simplify them into 10 clusters, as shown in Figure 4b. Clusters f1, f2, f3, f4 and f5 are forward plays, and b1, b2, b3, b4 and b5 are backward plays.

4.1 Covariance in Bayesian Networks

Although ball's landing position is not accurate enough to support a clustering decision, we can attach a probability to each possible cluster and find an interpretation which gives an optimal global probability. Using Bayesian rule, the probability of landing position λ belongs to cluster i can be updated from step k to $k + 1$ as following:

$$p_i^{(k+1)}(\lambda) = \frac{p_i^{(k)}(\lambda) [1 + q_i^{(k)}(\lambda)]}{\sum_{\lambda} p_i^{(k)}(\lambda) [1 + q_i^{(k)}(\lambda)]} \tag{4}$$

where $q_i^{(k)}(\lambda)$ is an updating factor:

$$\begin{aligned} q_i^{(k)}(\lambda) &= \frac{1}{n} \sum_j \sum_{\lambda'} r_{ij}(\lambda, \lambda') p_i^{(k)}(\lambda') \\ r_{ij}(\lambda, \lambda') &= \frac{\text{cov}_{ij}(\lambda, \lambda')}{\sigma_i(\lambda) \sigma_j(\lambda')} \\ \text{cov}_{ij}(\lambda, \lambda') &= p_{ij}(\lambda, \lambda') - p_i(\lambda) p_j(\lambda') \end{aligned} \tag{5}$$

Parameters λ and λ' are two consecutive landing positions. From the statistics of 58 patterns, we have the covariance matrix $\text{cov}_{ij}(\lambda, \lambda')$ as following:

$$\begin{bmatrix} -.014 & -.006 & -.006 & .011 & -.006 & 0 & .011 & .002 & .032 \\ .002 & .001 & -.006 & -.006 & .002 & .002 & -.006 & -.005 & 0 \\ 0 & .012 & 0 & .003 & .012 & .003 & .012 & -.005 & -.008 \\ .002 & -.020 & -.037 & .005 & .039 & -.037 & -.037 & -.011 & -.009 \\ -.008 & 0 & 0 & .034 & -.008 & -.017 & -.008 & .028 & -.002 \\ .003 & .012 & .003 & 0 & .012 & 0 & .012 & -.004 & .011 \\ .027 & .010 & .001 & -.016 & .010 & .001 & -.007 & 0 & -.013 \\ .017 & -.007 & .004 & .002 & -.001 & -.003 & -.005 & -.003 & -.004 \\ -.017 & .012 & .021 & -.013 & -.011 & .002 & .016 & -.005 & -.009 \end{bmatrix}$$

Formulae (4) and (5) have defined a simple Bayesian network, and can be applied to both landing identification and trajectory classification. However, in tennis tactical analysis, both landing location and trajectory classification are correlated. We define a two-level Bayesian network. The low level of the network works on landing location and trajectory classification separately and the high level network combines their probabilities. Both the low level and the high level work in tandem.

4.2 Feature Attributes of the Patterns

Our intelligent agent works on the feature attributes of each pattern. Following are feature attributes of our system:

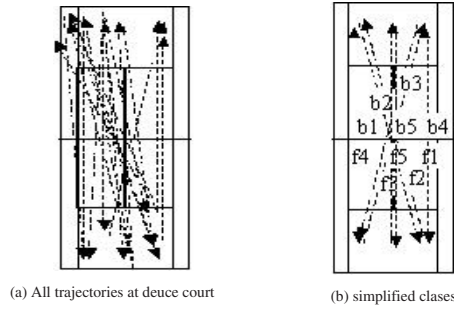


Fig. 4. Trajectory classification

- T_is_(x): T, F: $x \in \{f1, f2, f3, f4, f5, b1, b2, b3, b4, b5\}$
- B_is_(y): T, F: $y \in \{1..9\}$

Based on these attributes, the tactic pattern in Figure 1 can be represented as:

- T_is_f3 = T;
- B_is_7 = T;
- T_is_b3 = T;
- B_is_5 = T;
- T_is_f4 = T;
- B_is_1 = T;
- B_is_x = F: $x \in \{1..9\}$;

The rest of patterns can be represented in the same way. The total number of attributes is 332. The search space is 2^{332} . It is impossible to use exhaust search. The solution is obtained by induction using C4.5 [10].

4.3 Classification Using Inference

We can obtain Bayesian probability of ball landing clusters. However, obtaining a quantitative measure of the classification of trajectory is difficult. Even if we obtained a quantitative measure of trajectories, they were in the different dimensions from the probability of ball landing. We cannot simply add them up in calculation. To solve this problem, we employ an intelligent agent.

A generic entry in the network built in Section 4.2 is the probability of a conjunction of particular assignments to each pattern, such as $P(X_1 = e_1 \wedge \dots \wedge X_n = e_n)$. We use the notation $P(e_1, \dots, e_n)$ as an abbreviation for this. The value of this entry is given by:

$$P(E) = P(e_1, \dots, e_n) = \prod_{i=1}^n P(e_i | Parents(X_i))$$

We use E_x^+ to represent all e_i that are connected to X through its parents, and E_x^- to represent all e_j that are connected to X through its children. To inference the probability of X under evidence E , $P(X|E)$, we have an algorithm:

- $P(X|E) = P(X|E_x^-, E_x^+)$;
- compute the contribution of $P(X|E_X^+) = \sum_e P(X|e) \prod_i P(e_i|Parents(e_i))$
- compute $P(E_X^-|X) = \beta \prod_i \sum_{y_i} P(E_{Y_i}^-|y_i) \sum_{z_i} P(y_i|X, z_i) \prod_j P(z_{ij}|Parents(z_{ij}))$
- from $P(X|E) = \alpha P(E_X^-|X) P(X|E_X^+)$, compute

$$P(X|E) = \alpha P(E_X^-|X) \sum_e P(X|e) \prod_i P(e_i|Parents(e_i))$$

The computation involves recursive calls that spread out from X along all paths in the network. The recursion terminates on evidence nodes, root nodes and leaf nodes. Each recursive call excludes the node from which it is called, so each node in the tree is covered only once. Hence the algorithm is linear in the number of nodes in the network [12].

5 Empirical Results

The pattern classification process is very much a pattern identification process. The algorithm defined in Section 4 has to be implemented for all 58 patterns. When we have a sequence of playing, we have to detect the existences of any patterns. A separate algorithm will be needed to decide the threshold of pattern existences for the index purpose. This section provides an experiment on the implementation of the pattern shown in Figure 1. We test the implementation using a video sequence shown in Fig. 5.



(a) serve to the far end



(b) return to the serve



(c) return to the open court to win the point

Fig. 5. Trajectory of serving wide to open the court pattern

In the video sequence, there are three serve and returns. We initialize the probability of each hit to an equal value, ie, 1/10. After 25 iterations, we obtain:

	f1	f2	f3	f4	f5	b1	b2	b3	b4	b5
r1	0.09	0.04	0.61	0.10	0.14	0.003	0.004	0.003	0.008	0.002
r2	0.004	0.005	0.007	0.005	0.004	0.11	0.10	0.445	0.12	0.20
r3	0.04	0.036	0.19	0.494	0.22	0.005	0.004	0.003	0.005	0.003

The proper assignment of each return is labeled in bold figures, ie, return 1 is *f3*, return 2 is *b3* and return 3 is *f4*. The pattern can be classified as “serve wide to open the court”.

6 Conclusion

We develop a new scheme to summarize tennis video for training purpose. The scheme combines quantitative feature with binary attributes in classification using an intelligent agent. We have manually test it in a small number of patterns and the results are promising. The large scale testing is yet to be carried out.

Acknowledgement. This project is supported by Australia Research Council Linkage Grant (LP0347156).

References

1. BBC. http://www.bbc.co.uk/pressoffice/pressreleases/stories/2003/06_june/10/hawk_eye.shtml, 06 2003.
2. DSI. <http://www.moit.gov.il/root/hidden/ipc/advantages-stories1.html#dsi>, 2003.
3. A. Ekin and A. M. Tekalp. Automatic soccer video analysis and summarization. <http://www.ece.rochester.edu/~ekin/papers/CPapers/spie2003.pdf>, 2003.
4. Hawk-Eye. <http://news.bbc.co.uk/sport1/hi/tennis/2977068.stm>, 06 2003.
5. J. S. Jin. Computational simulation of edpth perception in the human visual system. In *Proceeding on 16th Annual Conference of the Cognitive Science Society, Georgia*, pages 451–456, 1994.
6. J. S. Jin, Z. Zhu, and G. Y. Xu. A stable vision system for moving vehicles. *IEEE Intelligent Transportation Systems*, 1(1):32–39, 2000.
7. B. Li and M. I. Sezan. Event detection and summarization in american football broadcast video. In *Proceeding on SPIE Storage and Retrieval for Media Databases*, volume 4676, pages 202–213, 2002.
8. G. S. Pingali, Y. Jean, and I. Carlbom. Real time tracking for enhanced tennis broadcasts. In *Proceeding on IEEE Comp. Vision and Patt. Rec. (CVPR)*, pages 260–265, 1998.
9. QUESTEC. <http://www.questec.com/q2001/news/1999/030599.html>, 2003.
10. J. R. Quinlan. *Programs for Machine Learning*, chapter 4. CA: Morgan Kaufmann, 1993.
11. Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proceeding on ACM Multimedia*, pages 105–115, 2000.
12. S. J. Russell and P. Norvig. *Artificial Intelligence - a Moden Approach*. London: Prentice-Hall, 1995.
13. USTA. *Tennis Tactics - Winning Patterns of Play*. Champaign: Human Kinetics, 1996.
14. Z. Zhang. *A Flexible New Technique for Camera Calibration*. Microsoft Research Microsoft Corporation, Redmond, technical report msr-tr-98-71 edition, 1999.
15. W. Zhou, A. Vellaikal, and C. C. J. Kuo. Rule-based video classification system for basketball video indexing. In *Proceeding on ACM Multimedia Workshops*, pages 213–216, 2000.

Performance Evaluation of High-Speed TCP Protocols with Pacing

Young-Soo Choi, Kang-Won Lee, and You-Ze Cho

School of Electrical Engineering and Computer Science
Kyungpook National University, Korea
{yschoi, kw0314}@eeecs.knu.ac.kr, yzcho@ee.knu.ac.kr

Abstract. The congestion control mechanisms of the current standard TCP can encounter problems in high-speed wide area networks due to its slow response with a large congestion window. Several congestion control proposals have already been suggested to solve this problem and mainly consider two properties: TCP friendliness and bandwidth scalability, to ensure that a protocol does not take away too much bandwidth from TCP, while utilizing the full bandwidth of high speed networks. However, further studies on the TCP friendliness of high-speed TCP are still needed. Recent studies have pointed out that existing schemes have a severe RTT unfairness problem, where competing flows with different RTTs can consume considerable unfair bandwidth shares. Burstiness is one of the main reasons behind such problems. As the congestion window achieved by a high-speed TCP connection can be quite large, there is a strong possibility that the sender will transmit a large burst of packets. As such, the current congestion control mechanisms of high-speed TCP can lead to bursty traffic flows in high speed networks, with a negative impact on both TCP friendliness and RTT unfairness. The proposed solution to these problems is to evenly space, or pace packets sent into the network over an entire round-trip time, so that packets are not sent in a burst. Accordingly, the current paper evaluates this approach with a high bandwidth-delay product network and shows that pacing offers better TCP friendliness and RTT fairness without degrading the bandwidth scalability.

1 Introduction

TCP has already been widely adopted as a data transfer protocol for the Internet. However, it has been reported that, as the bandwidth-delay product continues to grow, TCP substantially underutilizes the network bandwidth and will eventually become a performance bottleneck itself [1]-[5]. For example, according to [2], for TCP to increase its window to a full utilization of 10Gbps with 1500-byte packets, this will require over 83,333 RTTs. With 100ms RTT, this would take approximately 1.5 hours, and for full utilization in a steady state, the loss rate cannot be more than 1 loss event per 5,000,000,000 packets, which is less than the theoretical limit of the network's bit error rate. Consequently, it is impossible to achieve such a large throughput with TCP, mainly because TCP decreases its

congestion window too drastically when packet losses occur, yet only increases it very slightly when no packet loss is experienced.

Recently, various adaptive schemes have been designed that offer more flexibility, wider bandwidth scalability, and fairer competition with standard TCP. Such schemes include High Speed TCP (HSTCP) [2], Scalable TCP (STCP) [3], FAST [4], eXplicit Control Protocol (XCP) [1], and Binary Increase TCP (BI-TCP) [5]. XCP generalizes the Explicit Congestion Notification (ECN), enabling more (or explicit) information to be sent about the degree of congestion in the network. XCP gives predominant efficiency, fairness, and stability. However, since XCP requires XCP senders, routers, and receivers to be deployed, deployment issues still remain. In this paper, the above-mentioned protocols such as HSTCP, STCP, and BI-TCP are referred to as high-speed TCP protocols.

As the congestion window achieved by an existing high-speed TCP flow can be quite large, there is a strong possibility that the sender may send a large burst of packets in response to a single acknowledgement. Since the bursty behavior of high-speed TCP can lead to bursty traffic flows in high speed networks, with a considerable negative impact on TCP friendliness and RTT unfairness, existing high-speed TCP schemes need to include a means limiting bursts. A proposed solution is to pace packets sent into the networks over an entire round-trip time to avoid packets being sent in a burst. The use of pacing also enables the limiting of burstiness to be decoupled from congestion control, thereby allowing more flexible protocol design for congestion control. Therefore, the current paper conducts a quantitative evaluation of pacing with the most promising high-speed TCP schemes.

The remainder of this paper is organized as follows: Section II briefly discusses some high-speed TCP congestion control algorithms and pacing, then section III describes burstiness problem of high-speed TCP. Section IV describes the implementation of pacing and the simulation configuration. Section V presents the simulation results and some final conclusions are given in Section VI.

2 Related Work

The importance of congestion control is now widely acknowledged and extensive work has already been done to enhance the performance of TCP. TCP congestion control is composed of two major algorithms: slow-start and congestion avoidance algorithms. TCP uses a variable called congestion window (*cwnd*) and cannot inject more than *cwnd* segments of unacknowledged data into the network. The TCP congestion avoidance algorithm is called AIMD (Additive Increase Multiplicative Decrease) and is the basis for steady state congestion control. In the congestion avoidance phase, TCP increases the *cwnd* by one packet each RTT and reduces the *cwnd* by half in the event of a packet loss.

2.1 High-Speed TCP Protocols

HSTCP was introduced by Floyd in [2] as a modification of the TCP congestion control mechanism to improve the performance of TCP in fast, long delay

networks. As such, HSTCP is designed to have a different response in an environment with a very low congestion event rate, and have the regular TCP response in an environment with a packet loss rate of at most 10^{-3} . HSTCP introduces a new relation between the average congestion window w and the steady-state packet drop rate p . For simplicity, this new HSTCP response function maintains the property that the response function gives a straight line on a log-log scale. The HSTCP response function is specified using three parameters: *Low_Window*, *High_Window*, and *High_P*. *Low_Window* is used to establish a point of transition and ensure TCP friendliness. The HSTCP response function uses the same response function as regular TCP when the current *cwnd* is at most *Low_Window*, and uses the HSTCP response function when the current congestion window is greater than *Low_Window*. Meanwhile, *High_Window* and *High_P* are used to specify the upper end of the HSTCP response function, where *High_P* is the specific drop rate needed in the HSTCP response function to achieve a *High_Window* as the average congestion window. The HSTCP response function is represented by new additive increase and multiplicative decrease parameters. These parameters modify both the increase and decrease parameters according to the *cwnd*.

STCP was described by Kelly in [3] and changes the traditional TCP Reno congestion control algorithm. Instead of using an additive increase, the increase is exponential and the multiplicative decrease factor b is set to 0.125. That is, the congestion avoidance algorithm of STCP is MIMD (Multiplicative Increase and Multiplicative Decrease).

In [5], Xu *et al.* revealed that notwithstanding their scalability and TCP friendliness properties, HSTCP and STCP would appear to have a serious RTT unfairness problem when multiple flows with different RTT delays are competing. And they introduced BI-TCP that attempts to correct the RTT unfairness. The protocol uses an additive increase and a binary search increase. When the congestion window is large, an additive increase with a large increment (32 packets per RTT) ensures linear RTT fairness as well as scalability. Whereas, with small congestion windows, a binary search increase is designed to provide TCP friendliness.

The congestion avoidance algorithms of TCP, HSTCP, STCP, and BI-TCP are briefly expressed in Table I. For more details, see [2], [3], and [5], respectively. Since most existing schemes concentrate on bandwidth scalability and TCP friendliness, the performance of high-speed TCP and the impact of its use on the present implementation of TCP have been highlighted. As such, the results show that existing high-speed TCP schemes can relieve bandwidth scalability to a certain extent. However, further studies in TCP friendliness and RTT fairness are still needed.

2.2 Pacing

One of the most widely used mechanisms for smoothing out TCP traffic is pacing. Pacing is a hybrid between rate control and TCP's use of acknowledgments to trigger new packet to be sent into the network. TCP pacing can be implemented

Table 1. TCP Congestion Control in Congestion Avoidance

Regular TCP	ACK: $w \leftarrow w + a/w$ LOSS: $w \leftarrow w - b \times w$	$a = 1, b = 0.5$
HSTCP	ACK: $w \leftarrow w + a(w)/w$ LOSS: $w \leftarrow w - b(w) \times w$	$if\ w > Low_Window(= 38)$ $a(w) = 0.1578 \times w^{0.8024} \times b(w)/(2 - b(w)),$ $b(w) = 0.052 \ln w + 0.6892$ <i>else</i> regular TCP congestion control
STCP	ACK: $w \leftarrow w + a \times w$ LOSS: $w \leftarrow w - b \times w$	$if\ w > Low_Window(= 16)$ $a = 0.01, b = 0.125$ <i>else</i> regular TCP congestion control
BI-TCP	ACK: $if(target_win < S_{max})$ $w \leftarrow w + (target_win - w)/w$ <i>else</i> $w \leftarrow w + S_{max}/w$ LOSS: $w \leftarrow w - b \times w$	$if\ w > Low_Window(= 14)$ $S_{max} = 32, b = 0.125$ <i>else</i> regular TCP congestion control

at either the sender or the receiver side. That is, pacing is accomplished at the sender (receiver) if, instead of transmitting a packet (ACK) every time an ACK (packet) is received, it is delayed to maintain the proper spacing between two successive packets (ACKs). Pacing, which evenly spaces a window of packets over the round-trip time, was first proposed in [7]. Since then, pacing has been applied to correct the compression of acknowledgments due to cross traffic, to avoid the slow start of TCP at the beginning of a connection, after a packet loss, or when an idle connection resumes. Similarly, pacing can be used to avoid burstiness in asymmetric networks caused by batching acknowledgments. More recently, using pacing across the entire lifetime of a flow has been suggested. For more details on these and other environments where pacing has been used, the reader is referred to [8] and the references therein.

3 Pacing for Improving Fairness and TCP Friendliness in High-Speed Networks

As the congestion window achieved by an existing high-speed TCP flow can be quite large, there is a strong possibility that the sender may send a large burst of packets in response to a single acknowledgement. One of the main reasons behind this burstiness is the phenomenon of ACK compression, which occurs when a bunch of ACKs are sent by the receiver over a longer interval, yet arrive at the sender during a smaller interval. This burstiness problem can also happen when there is congestion or reordering on the reverse path and the sender receives an acknowledgement acknowledging hundreds or thousands of new packets. In this case, high-speed TCP flows fill up the pipe, potentially resulting in synchronized loss, where packet loss occurs across multiple competing flows simultaneously. Further, since above bursty behavior allows few flows (particularly short RTT

high-speed TCP flows) to monopolize the network buffers, burstiness can result in a considerable negative impact on TCP friendliness and fairness.

Therefore, existing high-speed TCP schemes need to include a means limiting bursts. Although mechanisms for limiting burstiness of high-speed TCP are important, they have not been studied yet. In this paper, pacing is applied over an entire round-trip time to avoid data being sent in a burst. The general purpose of this paper was to study the effectiveness of pacing in high speed long distance networks as a mechanism for limiting burstiness. Of particular concern was the study of TCP friendliness and RTT fairness. The current paper evaluates the effectiveness of pacing with HSTCP, STCP, and BI-TCP. FAST and XCP are not considered, as XCP has deployment issues, while FAST is a delay based congestion avoidance (DCA) algorithm. The correlation between increased delays (or RTTs) and congestive losses has recently been challenged [6], thereby raising serious doubts as to the effectiveness of DCA algorithms given that their main assumption is that RTT measurements can be used to predict and avoid network congestion.

4 Simulation Configuration

Pacing was implemented into the ns [9] simulation code for TCP SACK. The pacing implementation used a variant of the leaky bucket algorithm and pacing was used throughout the lifetime of a flow. Timeouts were scheduled at regular intervals $rtt/cwnd$, and a packet transmitted from the window whenever the timer expired. As new data was acknowledged (altering the window size), the duration of the current and subsequent intervals was altered to adjust to the new rate. The TCP timestamp option was used to obtain accurate RTT samples, and fine grained timers used to send data at the appropriate rate. Regular TCP uses TCP SACK without pacing.

The topology used for the simulation experiments is shown in Fig. 1. Various bottleneck capacities and delays were tested. The bottleneck router used FIFO scheduling and a drop tail buffer management scheme. By default, the buffer size at the bottleneck router was set to 100% of the bandwidth-delay products of the bottleneck link, unless otherwise specified. To reduce the phase effect and synchronized feedback, a significant amount of background traffic was used in both directions, along with randomized RTTs and starting times. All the high-speed TCP flows used the forward direction. Also, for background traffic, web traffic (20% up to 50% of the bottleneck bandwidth when no other flows were present), twenty-five small TCP flows with a limited congestion window size under 64, and 2 to 4 long lived TCP flows were created in both directions for all the simulation.

5 Simulation Results and Discussion

This section compares the performance of the paced high-speed TCP and non-paced high-speed TCP through simulation. The main concern of the current

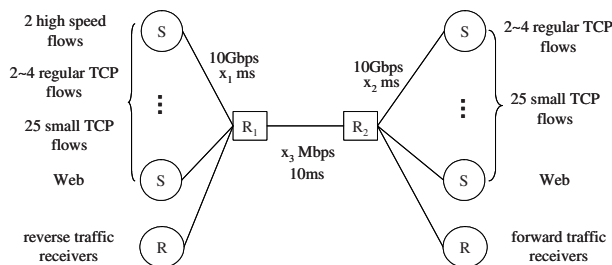


Fig. 1. The network topology for simulation.

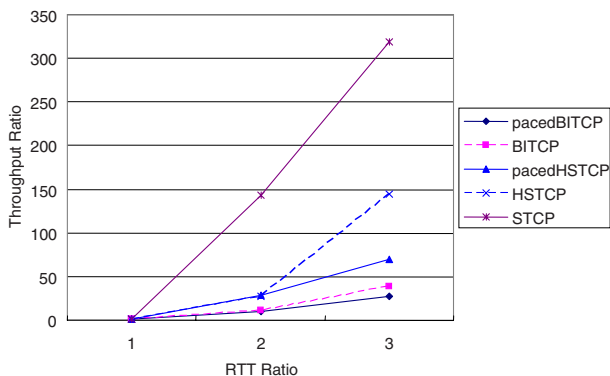


Fig. 2. The throughput ratio between two high-speed flows for different RTT ratios.

study was the effect of pacing on TCP friendliness and RTT fairness. The bandwidth scalability (i.e. link utilization) was omitted from the simulation results, because all protocols (with or without pacing) provide an approximately 100% link utilization.

5.1 RTT Fairness

In the experiment, two high-speed flows with a different RTT were used. The RTT of flow 1 was 40ms, while the RTT of flow 2 was varied among 40ms, 120ms, and 240ms. The bottleneck bandwidth was 2.5Gbps. To explore the impact of pacing on RTT fairness, the throughput of the high-speed flows was compared with and without pacing. Fig. 2 depicts the averaged throughput ratio between the two high speed flows with and without the presence of pacing. For example, when pacing was used with HSTCP, the throughput ratio was about 70, whereas without pacing it was 144 for RTT ratio 6. This is because pacing prevented the high speed flow with small RTT from overflowing the buffer. Similarly, the throughput ratio was reduced by half when the buffer size was set at 25% and 50% of the bandwidth-delay products for RTT ratio 6. But we did not present

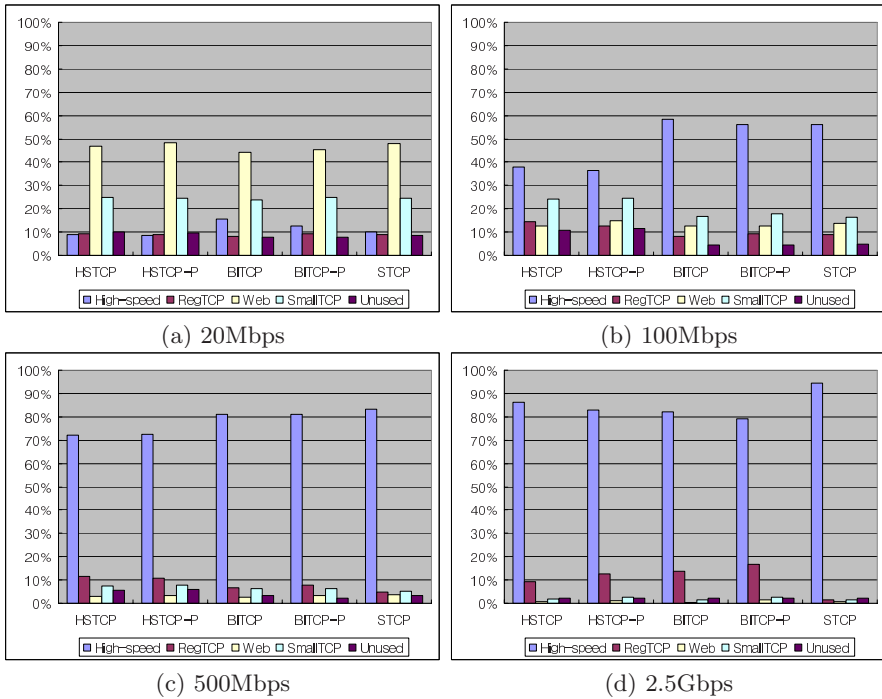


Fig. 3. TCP friendliness for various bottleneck bandwidths.

simulation results with various buffer sizes here due to lack of space. Therefore, the RTT fairness of the paced high-speed flows was much better than that of the non-paced high-speed flows, regardless of the type of high-speed TCP congestion avoidance algorithms.

5.2 TCP Friendliness

Four independent simulations were conducted with a different type of high-speed flows and bottleneck bandwidths. In every run, the same number of flows was used. Fig. 3 shows the percentage of the bandwidth shared by each flow type. For 20Mbps and 100Mbps, STCP and BI-TCP showed a similar TCP friendliness, while HSTCP showed a better TCP friendliness. For 500Mbps and 2.5Gbps, the share of bandwidth taken by the web, small TCP, and long-lived TCP flows was substantially reduced due to the TCP bandwidth scalability problem. Under 2.5Gbps, STCP was most aggressive, followed by HSTCP. BI-TCP became friendlier to TCP. The TCP friendliness of the paced HSTCP was comparable to that of the non-paced BI-TCP. Note that in every runs, high-speed flows with pacing used less bandwidth than non-paced high-speed flows. However, this reduced bandwidth was used by the background (small TCP, regular TCP, and web) traffic. And the amount of unused bandwidth both with and without

spacing scenarios is approximately the same. That is, paced high speed flows became friendlier to TCP. Accordingly, the paced high-speed TCP provided an enhanced TCP friendliness, regardless of the bandwidth and their congestion avoidance algorithms. Thus, pacing improved TCP friendliness relative to non-paced high-speed flows for all bandwidths without degrading the link utilization.

6 Conclusion

Although existing high-speed TCP schemes solve bandwidth scalability, there are still serious problems with RTT fairness and TCP friendliness. Accordingly, to solve these problems and limit the burstiness of high-speed TCP, the use of pacing with high-speed TCP was proposed. Simulation results showed that the paced high-speed TCP outperformed the non-paced TCP in terms of RTT fairness and friendliness without degrading the overall link utilization.

Acknowledgement. This work was in part supported by the ITRC, Ministry of Information and Communication Grant and the BK21 project.

References

1. D. Katabi, M. Handley, and C. Rohrs, "Internet Congestion Control for High Bandwidth-Delay Product Networks," In *Proceedings of the ACM SIGCOMM*, 2002.
2. S. Floyd, "HighSpeed TCP for Large Congestion Windows," *RFC3649*, 2003.
3. T. Kelly, "Scalable TCP: Improving Performance in Highspeed Wide Area Networks," *ACM SIGCOMM Computer Communication Review*, vol.33, pp. 83-91, 2003.
4. C. Jin, D. X. Wei and S. H. Low, "FAST TCP: motivation, architecture, algorithms, performance," In *Proceedings of the IEEE INFOCOM*, 2004.
5. L. Xu, K. Harfoush, and I. Rhee, "Binary Increase Congestion Control for Fast, Long Distance Networks," In *Proceedings of IEEE INFOCOM*, 2004.
6. J. Martin, A. Nilsson, and I. Rhee, "Delay Based Congestion Avoidance for TCP," *IEEE/ACM Transactions on Networking*, vol. 11, no. 3, pp. 356-369, 2003.
7. L. Zhang, S. Shenker, and David D. Clark, "Observations on the Dynamics of a Congestion Control Algorithm: The Effects of Two Way Traffic," In *Proceedings of the ACM SIGCOMM*, pp. 133-147, 1991.
8. A. Aggarwal, S. Savage, and T. Anderson, "Understanding the performance of TCP pacing," In *Proceedings of IEEE INFOCOM*, vol. 3, pp. 1157-1165, 2000.
9. The Network Simulator ns2, <http://www.isi.edu/nsnam/ns/>

Time-Triggered and Message-Triggered Object Architecture for Distributed Real-Time Multimedia Services

Doo-Hyun Kim, Eun Hwan Jo, and Moon Hae Kim

Konkuk University, Seoul, Korea
{doohyun, ehjo, mhkim}@konkuk.ac.kr
<http://kkucc.konkuk.ac.kr/~doohyun>

Abstract. In this paper, we present a new framework, called MMStream TMO, to effectively support the development of distributed multimedia applications using a real-time object model named the Time-triggered Message-triggered Object (TMO). The time-triggered spontaneous feature of TMO is used as a regulator against the irregular deliveries of media units caused by QoS non-guaranteed systems and communication channels. The message-triggered feature of TMO is used as a vehicle for delivering and broadcasting commands and control messages between MMStream TMOs. The global-time feature of TMO is utilized for facilitating measurement of network delay and delay jitter, and re-synchronization of intra-stream and inter-streams. The experiment preliminarily conducted on a single H.261 video stream showed that our scheme can tolerate and recover the network and end-system delay jitters effectively. This global time based de-jittering capability is expected to contribute to overall enhancement of the inter-stream and lip-synchronization accuracies.

1 Introduction

Today, multimedia is an important field of application for computers, wireless devices and embedded systems. The goal of a distributed multimedia application is to provide reliable high-quality multimedia services to users on any network. Common distributed multimedia services use streaming techniques that process and transport digitized media on a network. But, the current multimedia streaming technology is not suited to the development of a complex real-time multimedia application such as video conferencing system [1].

Firstly, most multimedia streaming technologies cannot guarantee reliable services to users. They do not provide timely service capabilities. Secondly, it is difficult to develop a distributed multimedia application that is executed on heterogeneous hardware platforms and operating systems. Thirdly, owing to the lack of real-time multimedia streaming APIs, it is not easy to develop complex multimedia applications.

In this paper, we present a real-time multimedia streaming architecture, called MMStream TMO, to facilitate the development of reliable multimedia

applications. Our architecture uses TMO model and execution engine as an underpinning[2,3]. The time-triggered spontaneous feature of TMO is used not only as a periodical media generator, but also as a regulator against the irregular deliveries of media units caused by QoS non-guaranteed systems and communication channels. The message-triggered feature of TMO is used as a vehicle for delivering and broadcasting commands and control messages between MMStream TMOs. With its APIs (MMTMOSL) and library (MMTMOSM), the MMStream TMO is expected to provide developers easy and consistent way for not easy and programmer-dependant issues such as transportation, transformation, and synchronization issues.

In section 2, we briefly describe the requirements for distributed multimedia streaming in real-time and the TMO scheme that is the basis of our approach. Section 3 presents the MMStream TMO framework. Section 4 presents a global-time based synchronization scheme and its experimental results. Finally, section 5 summarizes the paper with future works.

2 Backgrounds

2.1 Requirements for Multimedia Streaming

The streaming conducts the processing of a stream, that is, a sequence of media data units with temporal constraints. The processing of a stream is embodied by a chunk of codes or an object which receives a media data stream, changes certain characteristics of the stream by applying filtering operations, and outputs the processed stream. As a basic element of a multimedia application, typical functions of the stream processing can be highlighted as follows:

- *Transportation* of sequential multimedia streams from source such as file, device and network to sink such as file, device and network.
- *Transformation and processing* (filtering, encoding, decoding, multiplexing, de-multiplexing, mixing, and synchronization) of multimedia.
- *Synchronization* of audio and video so that those can be started and stopped at the same time and can be played at the same rate.

In addition to the above preliminary requirements, in order to provide and develop high quality multimedia streaming services, the following factors should be considered:

- Multimedia streams could contain large amounts of data, which must be processed in timely fashion.
- Data can come from many sources, including local files, computer networks and video cameras.
- A developer does not know in advance which hardware devices will be used at end-user's systems.

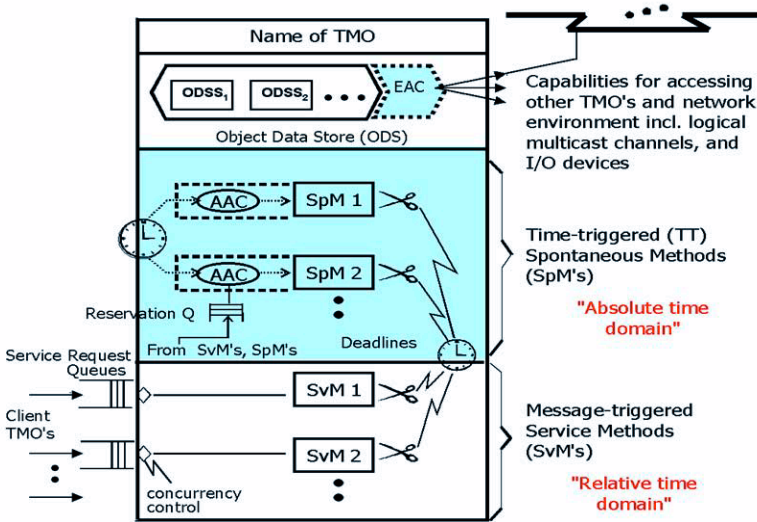


Fig. 1. Structure of TMO[2]

2.2 TMO Scheme

TMO is a natural, syntactically minor, and semantically powerful extension of the conventional object(s)[2,3]. Particularly, TMO is a high-level real-time computing object. Member functions (i.e., methods) are executed within specified time. Timing requirements are specified in natural intuitive forms with no esoteric styles imposed. As depicted in Fig. 1, the basic TMO structure consists of four parts:

- *Spontaneous Methods (SpM)*: a new type of method, also known as the time-triggered (TT) method. The SpM executions are triggered when the real-time clock reaches specific values determined at design time. A SpM has an AAC (Autonomous Activation Condition), which is a specification of the time-windows for execution of the SpM.
- *Service Method (SvM)*: conventional service methods. The SvM executions are triggered by service request messages from clients.
- *Object Data Store (ODS)*: the basic unit of storage which can be exclusively accessed by a certain TMO method execution at any given time or shared among concurrent executions of TMO methods (SpMs or SvMs).
- *Environment Access Capability (EAC)*: the list of entry points to remote object methods, logical communication channels, and I/O device interfaces.

2.3 Feasibilities of TMO for Modeling and Implementing Multimedia Streaming

Among the various features of the TMO programming scheme mentioned above, some of the fundamental features can enable efficient programming of complex

distributed multimedia applications. First, TMO uses a global time base[4] as an integral component and enables the programmer to design global time based coordination of distributed multimedia actions very easily. The scheme provides a sound foundation for programming and executing distributed multimedia actions requiring global synchronization amongst the media units or amongst the nodes collectively engaged in a multimedia presentation performance. Examples of such performance can be found when multiple video clips are played back at different nodes at exact starting time points predefined for each clip.

Secondly, the clear separation and BCC between SpMs and SvMs allows the use of the SpM, the time-triggered spontaneous method, as an accurate means for periodic stream processing. Third, the message-triggered method, SvM, can be used as an easy-to-program command and control channels during multimedia processing TMOs. Each multimedia processing object should have member functions for getting and sending commands, i.e., play, pause, and stop commands, as well as control data that are necessary for capability negotiations. It is also easy to establish a GUI (Graphic User Interface) program module through which the user can invoke SvMs of multimedia processing TMOs, especially those SvMs serving as receptionists for commands from the user. The member functions for command exchanges are necessarily not only for facilitating the control by the user, but also for the cooperative session control by the TMOs.

3 Multimedia Streaming Based on TMO

3.1 MMStream TMO

MMStream TMO is a special form of TMO for processing multimedia streams at application level. Distributed MMStream TMOs in multiple nodes cooperatively work to embody a multimedia streaming. The use of MMStream TMO is supported by MMTMOSL and TMOSL as depicted in Fig. 2. In order to make full use of the TMO scheme, the basic elements such as ODS, SpM, SvM and Gate are used as shown in Fig. 3.

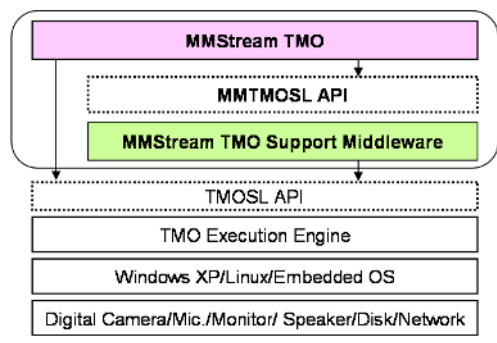


Fig. 2. Relationships among MMStream TMO, MMTMOSL API, MMStreamSM

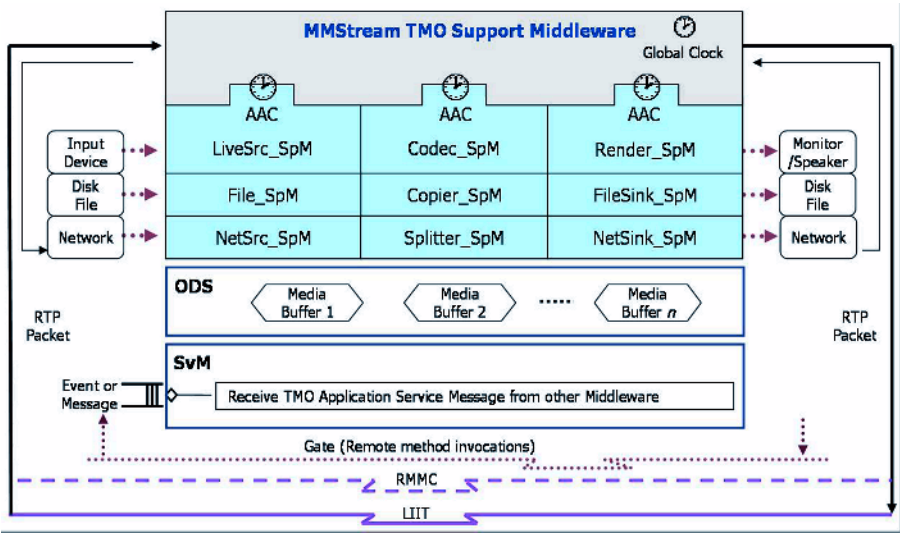


Fig. 3. MMStream TMO Architecture

- SpM supports to capture, transport, and display the stream data. That is, on or more SpMs in an MMStream TMO process a multimedia stream and control streaming flow. Each SpM uses ODS to move stream data from one SpM to another. With activation time specified in the AAC section of each SpM, inter- and intra-media synchronization can be controlled.
- SvM handles user's request messages. According to a user's request, a corresponding SvM performs play, pause, or stop operations, and also trigger SpM's dynamically.
- ODS is used to store stream data for buffering so as to synchronize intra-media stream data. Also, ODS is utilized as a connection point for SpM filters that take part in streaming. Access to ODS is performed by reading and writing operations of SpM's periodically.
- Gate is used as a communication channel for control messages by SvM, and the other fundamental way of connecting TMOs, RMMC(Real-Time Multicast and Memory-Replication Channels)[2] is used for broadcasting multimedia data units.

3.2 MMTMOSM(MMTMO Support Middleware) and MMTMOSL

MMTMOSM is to support synchronization abilities in real-time streaming services by providing session management, resource management, services scheduling and transparent transmission over heterogeneous networks. Currently, MMTMOSM has precise real-time capabilities by using TMO SM/Linux[5]. Meanwhile, MMTMOSL is a high-level application programming interfaces (APIs) abstract-

Table 1. Summary of MMStream Support Library Classes

Class names	Class descriptions
MMTMOSM	An abstraction for basic MMTMOSM services related to time-based media capturing, processing, and presentation. Programmers can register his/her own MMStream TMOs by instantiating MMTMOSM classes.
MMStreamTMO	A template that programmers inherit to define his/her own TMO class.
MediaSource	An abstraction for any object that receives data from a stream source. Once constructed, it functions automatically by a SpM.
MediaSink	An abstraction for any object that provides data to a stream destination. Once constructed, it functions automatically by a SpM.
MediaProcessor	An abstraction for any object that transforms media data including encoding, decoding, weaving and splitting.
MediaOds	An abstraction for any object that stores media data for buffering.
MediaTime	An abstraction of a global time for specifying a streaming time value such as start time, stop time, and current time.
MediaPlayer	A class for containing the application specific codes implementing use interface.

ing the MMTMOSM functionalities, and provides eight C++ classes as summarized in Table 1.

4 Synchronization Scheme and Experimental Results

A multimedia data stream transferred from a source node to one or more sink nodes consists of consecutive logical data units (LDUs). The data elements in a stream must be presented at the sink node in manners exhibiting the same temporal relationship that existed among the data elements when they are captured at the source node. One of the most commonly encountered situations where such requirement is evident is where the simultaneous playback of audio and video must be done with the so-called "lip synchronization." If both media are not played in good synchronization, the result will not be accepted to be of high quality.

In order to achieve synchronization, the global timing feature[5] provided by MMTMOSM can be used relatively easily by specifying temporal conditions in AAC for Source_SpM and Sink_SpM at design time as follows:

$$\begin{aligned}
 & \textit{for } T = \textit{from } TMO_START + S + F \\
 & \quad \textit{to } TMO_START + P \\
 & \quad \textit{every } SY \\
 & \quad \textit{start_during } (T, T + OS_DELAY_FOR_STREAMING) \\
 & \quad \textit{finish_by } T + DEADLINE_FOR_STREAMING)
 \end{aligned} \tag{1}$$

The *S* and *P* denote *Start* time and *StoP*time, and *SY* and *F* denote *Intra-SYN*chronization time and *Pre-Fe*tch time for initial buffering before streaming

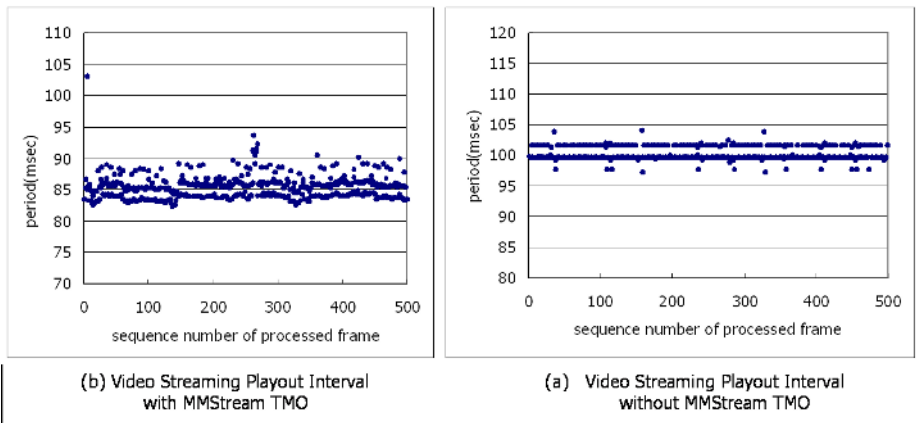


Fig. 4. Periodicity Comparisons of Video Playout Interval

is started. A SpM with the above AAC starts at the time $TMO_START + S + F$ and is executed until the time $TMO_START + P$ with the period SY . Also, the SpM must start between T and $T + OS_DELAY_FOR_STREAMING$ and complete its task within $T + DEADLINE_FOR_STREAMING$. T is a time variable and TMO_START refers to the start time of the TMO execution engine. $OS_DELAY_FOR_STREAMING$ is the time spent by OS for activating an SpM. Each stream may have different start time S . By start time of each stream, inter-synchronization times can be specified. Also SY represents intra-synchronization. Therefore, by giving appropriate values to these times, the flow of streaming can be controlled.

A well-synchronized global time base is essential for proper and timely operation of the distributed real-time multimedia system. Local clocks in distributed computing nodes will diverge due to their difference in clock drift rates. Each local clock should be adjusted periodically so that at any given time, the difference among local clocks of all participating distributed streaming nodes may be bounded within a specific deviation. TMO execution engine such as TMOSM uses the time information from a high-resolution performance counter in a local hardware. Although the main theme of this paper is to propose a novel framework using the TMO scheme as a sound foundation for building globally synchronized multimedia applications, we experimented a video streaming to show a preliminary effectiveness of using globally synchronized time-triggered streaming of MMStream TMO. While the Fig. 4(a) shows play-out intervals when we implemented a single video streaming without using MMStream TMO, the Fig. 4(b) does with MMStream TMO. The Fig. 4(b) shows that, when MMStream TMO is used, the deviation of the play-out interval becomes smaller. This means that the accuracy of lip synchronization and inter-stream synchronization will be easily conducted.

For the experiment, we used a local LAN and the H.261 video codec. We expect that the periodicity regulation power will be more effective within a reasonable range, if MMStream TMO is used for a WAN public Internet.

5 Conclusions

We proposed MMStream TMO based on the TMO scheme as a software framework for developing distributed real-time multimedia applications. The time-triggered spontaneous feature of TMO is used as a regulator against the irregular deliveries of media units caused by QoS non-guaranteed system and communication channels. The message-triggered feature of TMO is used as a vehicle for delivering and broadcasting command and control messages between MMStream TMOs. With its API (MMTMOSL) and Library (MMTMOSM), MMStream TMO is expected to provide developers easy and consistent way for complex and programmer-dependant issues such as transportation, transformation and synchronization issues. Currently, MMStream TMO has been implemented on TMOSM/XP and TMOSM/Linux [5]. In the near future, we focus on enriching the MMTMOSM library with diverse audio and video codecs, and producing viable application products such as on-line Quiz like *Wheel of Fortune(Jeopardy!)* of ABC-TV[6] where the globally synchronized multiparty audio-video conversations are mandated.

Acknowledgement. This research was supported by University IT Research Center Project, MIC, KOREA.

References

1. Sitaram, D. and Dan, A.: *Multimedia Servers: Application, Environments, and Design*. Morgan Kaufmann Publishers, San Francisco (2000)
2. Kim, K.H.: APIs for Real-Time Distributed Object Programming. *IEEE Computer*, (2000) 72–80
3. Kim, K.H., Ishida, M., and Liu, J.: An Efficient Middleware Architecture Supporting Time-Triggered Message-Triggered Objects and an NT-based Implementation. *ISORC*, (1999) 54–63
4. Kopetz, H.: *Real-Time Systems: Design Principles for Distributed Embedded Applications*. Kluwer Academic Pub., ISBN: 0-7923-9894-7, Boston (1997)
5. Kim, H.J., Park, S.H., Kim, J.G., and Kim, M.H.: TMO-Linux: A Linux-based Real-time Operating System Supporting Execution of TMOs, *ISORC* (2002)
6. <http://www.seeing-stars.com/ShowBiz/WheelOfFortune.shtml>

A Novel Multiple Time Scale Congestion Control Scheme for Multimedia Traffic*

Hao Yin, Chuang Lin, Ting Cui, Xiao-meng Huang, and Zhang-xi Tan

Computer Science Department, Tsinghua University, 100084, Beijing, China
{hyin, clin, tcui, xmhuang, zxtan}@csnet1.tsinghua.edu.cn

Abstract. A novel two time scale congestion control scheme is proposed in this paper, which is called Adaptive Wavelet and Probability-based scheme (AWP). On normal time scales of 20-200 ms, it is a smooth and TCP-friendly congestion control; on large time scales of 1-5 second, it detects persistent shifts in overall network contention and modulates the rate control exhibited on normal scale. AWP exploits self-similar property of network traffic and predicts future network available bandwidth to enhance performance, and it is adaptive to network environment changes through an auto-correction algorithm. With comparable experiments among TFRC, TCP and AWP over NS platform, AWP has been proved to be able to greatly improve the quality of service(QoS) of multimedia data transmission while avoiding congestion collapse on the network.

Keywords: Congestion Control, Multiple Time Scale, Multimedia Traffic, Self-similar.

1 Introduction

Recently, multimedia data transfer is one major type of traffic over the Internet. However, large amount of multimedia data on the Internet may cause serious network congestion. To resolve these problems, many congestion control schemes are proposed, either windows-based or rate-based [1]. Current real-time streaming applications on the Internet typically rely on rate-based congestion control, for such control can generate a smooth and TCP-friendly data flow such as TFRC [4]. Meanwhile, abundant studies have shown that network traffic exhibit self-similar properties statistically, which cause low quality of service, especially when the aggregated traffic is high [2]. So, in this paper we aim to apply the self-similar property of network traffic to designing a novel rate-based congestion control scheme in order to maximize the QoS of delivered multimedia data while preventing congestion collapse of the network.

* This work is supported by the National Natural Science Foundation of China (No. 60372019), China Postdoctoral Science Foundation (No.2003034152), the Projects of Development Plan of the State Key Fundamental Research (No.2003CB314804) and the Projects of Development Plan of the State High Technology Research (No.2001AA112080).

Park et al [2] proposed Selective Aggressiveness Control (SAC). SAC constructs a 2-level time scale congestion control protocol that works concurrently across two time scales. Although SAC has proved the effectiveness of applying self-similar properties to improving the congestion control performance, it does not take the demand of multimedia data delivery into account and is not theoretically modeled, which essentially expresses the disadvantages of this scheme. Ribeiro et al in [3] developed a model-based technique called Delphi algorithm. It gives accurate cross-traffic estimates for higher link-utilization level across an end-to-end path with a parsimonious multifractal parametric model, called the multifractal wavelet model (MWM); however it over-estimates the cross-traffic at lower link-utilization for the lack of adaptation.

In this paper, a novel multiple time scale congestion control scheme for multimedia traffic is proposed, which is called Adaptive Wavelet and Probability-based (AWP) scheme. AWP consults the framework of SAC, while it adds extended MWM (EMWM) and novel rate control algorithm into the framework. With a series of simulation experiments, AWP has the following unique advantages against TCP and TFRC: (1) AWP greatly improves the network bandwidth utility with low packet loss probability; (2) AWP contains good adaptation to various network environment; (3) AWP preserves smoothness and TCP friendliness of output rate, which are very important for the QoS improvement of multimedia transmission and stability of the network.

The rest of this paper is organized as follows: Section 2 presents an overview of AWP. Section 3 gives a detailed discussion on the crucial algorithms of AWP. The NS experiment results are illustrated in Section 4. At last, Section 5 closes this paper with conclusion.

2 Overview of AWP Congestion Control Scheme

Fig 1 shows the structure of AWP control scheme. AWP constructs two control levels. The first component - acting at 20-200 ms time scales - uses implicit prediction to affect rate control. The second component - acting at 1-5 s time

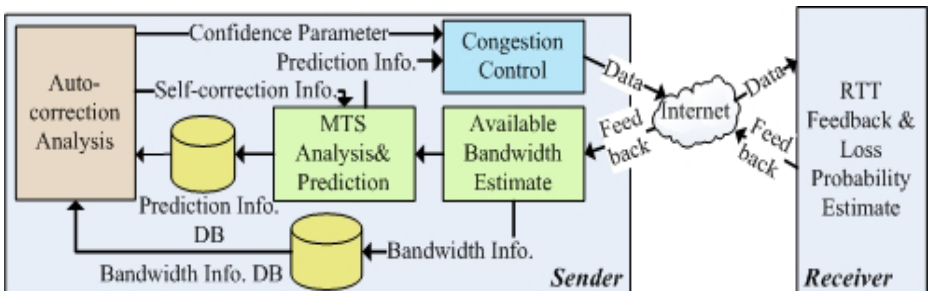


Fig. 1. A brief view of AWP framework. MTS is abbreviation for Multiple Time Scale, DB is abbreviation for Database

scales - uses explicit prediction to detect persistent shifts in overall network contention and modulates the rate control in the first component. Meanwhile AWP adopts auto-correction analysis to examine the validity of the predicted information, to improve the efficiency of prediction algorithm.

2.1 Sender Functionality

At first, the sender side collects the feedback information on normal time scale from the receiver side, and estimates the available bandwidth of end-to-end path using TCP throughput formula [4]:

$$A = \frac{s}{R\sqrt{\frac{2p}{3}} + t_{RTO}(3\sqrt{\frac{3p}{8}})p(1 + 32p^2)} \tag{1}$$

where the available bandwidth A (in *bytes/sec*) is a function of packet size s , round-trip time R , steady-stat loss event probability p , and the TCP retransmission timeout value t_{RTO} .

Then, A is imported into the Multiple Time Scale Analysis and Prediction (MTSAP) module, where EMWM is used to approximate current and past values of A and to predict future network traffic on large time scale. Meanwhile, volumes of A form a time series $A(t)$:

$$A(t), t = \omega, \dots, \omega + Ts, \omega \geq 0, Ts > 0 \tag{2}$$

where ω is the start time of i th control interval, and Ts is the approximation interval, which also expresses the large time scale. The time schedule of control intervals of the sender is shown in Fig 2.

AWP runs control intervals periodically. Within each control interval, AWP spends Ts in EMWM approximation, as well as δ in MTSAP prediction and congestion control ($\delta \ll Ts$). At time $\omega + Ts$ of i th control interval, approximating process of EMWM is finished, and MTSAP module generates a new time series $B(t)$, which is composed of the values of predicted available bandwidth for $i + 1$ th control interval:

$$B(t), t = \omega + Ts, \dots, \omega + 2Ts, \omega \geq 0, Ts > 0 \tag{3}$$

where ω and Ts are defined above. The congestion control module of i th control interval takes effect at time $\omega + Ts + \delta$, whereas the next EMWM approximation starts at time $\omega + Ts$. $A(t)$ is stored in Bandwidth Information Database for further use in auto-correction analysis. And $B(t)$ is stored in Prediction Information Database for the same purpose.

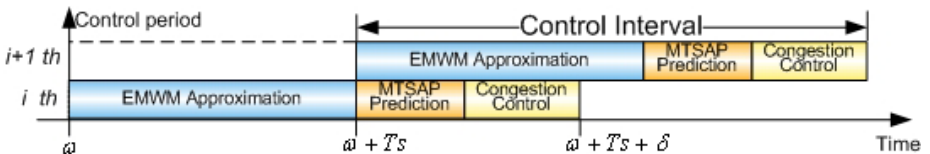


Fig. 2. Time schedule of control intervals of AWP

Adaptation of AWP is achieved by analyzing the series $A(t)$ and $B(t)$, and the result is used by congestion control module and MTSAP module, to guide congestion control algorithm and to auto-correct the prediction algorithm, respectively.

At time $\omega + Ts + \delta$ (shown in Fig 2), congestion control module makes the control decision on large time scale according to the following strategy:

$$\begin{cases} \text{if } \alpha > \theta \text{ then } \Lambda = E(B(t)) \\ \text{Else wait for next prediction} \end{cases} \quad (4)$$

where θ is a threshold to decide whether to trust the prediction information, Λ is the controlled throughput on large time scale for the next control interval, and $E(B(t))$ is the mean of $B(t)$. The choice of θ is discussed in Section 3.3. On normal time scale, the congestion control algorithm performs TCP-friendly and smooth rate control according to Λ . The detailed algorithm is also discussed in Section 3.3.

2.2 Receiver Functionality

The receiver side firstly estimates the packet loss event probability p . A loss event consists of several packet losses within a round-trip time. If the data flow sends N packets per round-trip time, and assume a Bernoulli loss model with loss probability p_{loss} , the loss event probability p , as a function of number of loss events per packet sent, is given by [4]:

$$p = \frac{1 - (1 - p_{\text{loss}})^N}{N} \quad (5)$$

Then the receiver side feeds back p to the sender together with the time stamp of the latest received packet provided for the sender to calculate the round-trip time (RTT).

3 Crucial Algorithms of AWP

3.1 Multiple Time Scale Prediction for Self-Similar Traffic

In this paper we propose EMWM to approximate the network traffic. EMWM is based on Haar wavelet system, and it can portray the self-similar property of network traffic properly. Fig 3 shows the binary tree of scaling coefficients [3] from normal time scale to large time scale of EMWM. As defined in Section 2.1, $A(t)$ is the estimated available bandwidth series. For simplicity, we assume $A(t)$ is composed of only two elements $A(1)$ and $A(2)$. At first, we apply wavelet transform [11] onto $A(t)$ using formula (6).

$$U_{j,k} = W(A(t)) = \int A(t)2^{j/2}\phi(2^j t - k) dt \quad (6)$$

where $\phi_{j,k}(t)$ is band pass wavelet function, and $U_{j,k}$ measures the local mean around time $2^{-j}k$. Thus $U_{j+2,4k}$ and $U_{j+2,4k+1}$ are the scaling coefficients of $A(1)$ and $A(2)$. In binary tree structure, the following formulas hold [3]:

$$U_{j,k} = U_{j+1,2k} + U_{j+1,2k+1} \quad (7)$$

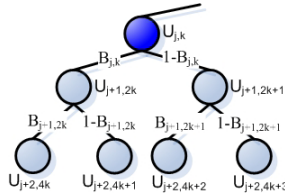


Fig. 3. Binary tree of scaling coefficients from normal scale to large scale

$$U_{j+1,2k} = B_{j,k} \times U_{j,k}, \quad U_{j+1,2k+1} = (1 - B_{j,k}) \times U_{j,k} \tag{8}$$

Formula (7) and (8) show that the coefficients on large time scale (parent node) can be obtained by the summary of those on normal time scale (children nodes), and the coefficients on normal time scale can be obtained by splitting the parent with a random multiplier $B_{j,k}$, where $B_{j,k}$ is distributed between 0 and 1. A general choice of $B_{j,k}$ is the symmetric Beta Distribution [10].

The coefficients of $A(t)$ on normal time scale are accumulated upward by formula (7) to obtain traffic load information on large time scale and $B_{j,k}$'s parameters at different scales [7]. Then the coefficient on large time scale is splitted downward and applied inverse wavelet transform to generate the prediction series $B(t)$. In Fig 3, $U_{j+2,4k+2}$ and $U_{j+2,4k+3}$ represent the predicted scaling coefficients of $B(t)$ in wavelet domain. And with inverse transform of (6), the elements of $B(t)$ are converted to time domain, which predict available bandwidth on normal time scale.

However, the prediction from EMWM is not valid all the time, and it is necessary to make corrections from history to adjust the prediction result, so as to keep the prediction error rate low. The coefficients on normal time scale are adjusted by referring confidence parameter and the last predicted information:

$$B(t+Ts) = \alpha \times W^{-1}(U_{j+2,4k+i}) + (1-\alpha) \times B(t), \quad t \in [\omega, \omega+Ts], \quad i \in [2^{j-1}, 2^j - 1] \tag{9}$$

$B(t + Ts)$ is the predicted available bandwidth series. $W^{-1}(U)$ is the inverse wavelet transform of U . And the rest symbols are defined in Section 2.

3.2 Auto-correction Analysis

AWP applies Bayes' Theorem [10] to carry through the auto-correction analysis. The Bayes' Theorem has the following form:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} = P(M) \frac{P(D|M)}{P(D)} \tag{10}$$

The prior probability $P(M)$ is determined by EMWM. In this paper, it is set to 0.8, for a high validity of the model. $P(D)$ is the probability that $A(t)$ exhibits second-order self-similarity. $P(D|M)$ is the probability that $B(t)$ insulates the estimated available bandwidth. Thus the confidence parameter $\alpha = P(M|D)$ can be obtained from formula (10). It explains how appropriate the traffic model is fit to the real network environment based on observed data.

3.3 Congestion Control Scheme

Threshold for Large Time Scale Rate Control Scheme. Currently, the threshold θ in formula (4) is 0.5, which means if $B(t)$ is at least 50% accurate, AWP accepts $B(t)$; otherwise the prediction is failed so AWP makes no change of rate control on large time scale, and waits for next valid prediction, in order to avoid fault control decisions.

Normal Time Scale Rate Control Scheme. On normal time scale, AWP employs one similar scheme out of the class of non-linear TCP compatible congestion control schemes called Binomial Congestion Control Schemes, which are well suited for real time streaming applications [5,6]. The scheme changes its throughput smoothly:

$$\lambda' = \begin{cases} \lambda + \delta & \text{increase rate} \\ \lambda - 0.6 \times \sqrt{\lambda/RTT} & \text{decrease rate} \end{cases} \quad (11)$$

where λ' is the new throughput on normal time scale, λ is the current throughput, δ is an additive linear increase, and RTT is the round-trip time.

4 Simulation Experiments

We have run several experiments via NS platform [9] to test AWP under self-similar condition, and have made comparisons with TFRC scheme for both AWP and TFRC are designed for multimedia traffic.

4.1 ns-2 Experiment Set-Up

We simulated a network environment where 4 concurrent connections are multiplexed over a shared bottle-neck network. Fig 4 shows the network topology. In the experiments, either TFRC or AWP is running at one time. On node 2, we have set up 16 UDP agents transferring files with Pareto distributed sizes, acting as infinite sources. Another 16 UDP agents with the same traffic are deployed to act as finite sources in the middle of simulation, to generate more severe burstiness of traffic. The Pareto distribution parameter is set as 1.05, 1.25, 1.65 and 1.85 for different distribution shapes. There is also an FTP source with TCP protocol transferring packets at a constant rate on node 3. A typical run lasted for 200 seconds, with traces collected at 0.3s granularity. The time scale axis in all the figures denotes large time scale values.

4.2 The Experiment Results

Average Bandwidth Utilities at Various Time Scales. The time scale ranges 0.5s to 5s, with an step interval of 0.5s. Fig 5(left) shows that AWP has the highest bandwidth utility of 87.47% on average, whereas TFRC has 81.93%, and TCP only reaches 76.69%. Also, AWP has exhibited a variation

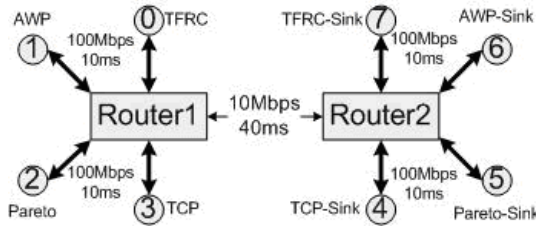


Fig. 4. Network topology for ns simulations with bottle-neck link

across different time scales, and it is because: first, TFRC and TCP do not control throughput on different time scales; and second, in AWP mechanism, different values of large time scale results in different response of controlled rate on normal time scale.

Dynamic Available Bandwidth Utility. The choice of the large time scale is 3 second, because at this point AWP has a higher link-utility with less packet loss. Fig 5(right) shows that AWP follows the available bandwidth more rapid and closely than TFRC at all times, which proves AWP more adaptive to various network conditions.

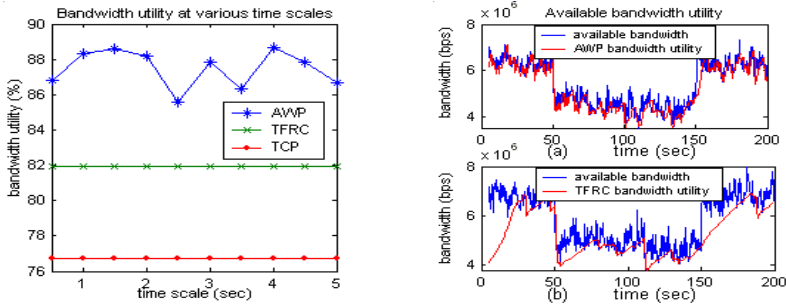


Fig. 5. LEFT: Average bandwidth utility at various scales of AWP, TFRC, and TCP; RIGHT: Dynamic available bandwidth utility of AWP and TFRC

Accuracy of Prediction. Fig 6(left) illustrates the prediction trace against the real one. Define the error rate:

$$Err_i = |p_i - r_i|/r_i \tag{12}$$

where p_i is the i th value of predicted available bandwidth, and r_i is the corresponding estimated value. Through calculation, the mean error rate is 0.0868, which means the prediction process can predict the future self-similar network traffic with high accuracy.

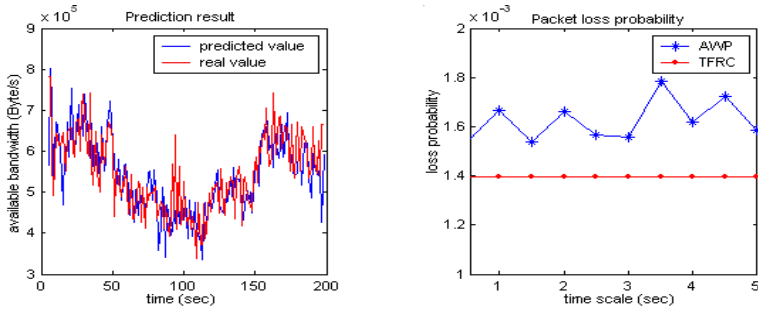


Fig. 6. LEFT: A comparison of predicted traffic and real Traffic; RIGHT: Packet loss probability of AWP and TFRC

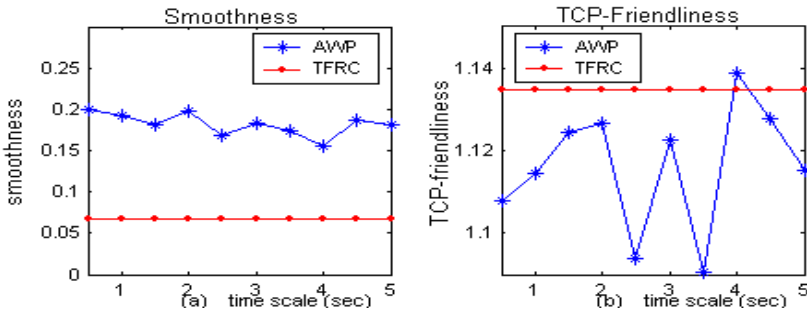


Fig. 7. AWP and TFRC smoothness and TCP-friendliness comparison

Packet Loss Probability. In Fig 6(right), AWP endures about 2 more packet loss events out of 10,000 ones than TFRC on average; however, the malfesance caused by such few drops can be eliminated by error control mechanism for multimedia. So AWP is as good as TFRC on packet loss performance.

Transfer Rate Smoothness. Let R_{tcp}^j , R_{awp}^j , R_{tfrc}^j denote the j th sampled transfer rate for TCP, AWP, and TFRC respectively, and N_{tcp} , N_{awp} , N_{tfrc} the total number of samples. Assuming the smoothness of TCP is 1, define

$$S_{awp} = \frac{\sum_{j=1}^{N_{awp}} |R_{awp}^j - R_{tfrc}^{j-1}|}{\sum_{j=1}^{N_{tcp}} |R_{tcp}^j - R_{tcp}^{j-1}|} \quad S_{tfrc} = \frac{\sum_{j=1}^{N_{tfrc}} |R_{tfrc}^j - R_{tfrc}^{j-1}|}{\sum_{j=1}^{N_{tcp}} |R_{tcp}^j - R_{tcp}^{j-1}|} \quad (13)$$

to express the smoothness of AWP and TFRC against TCP [8].

A lower result of formula (13) expresses a better smoothness of the mechanism. In Fig 7(a), AWP exhibits less smooth than TFRC, for a higher utility of available bandwidth, resulting in more changes of throughput. However, AWP is still much smoother than TCP, which guarantees AWP suitable for multimedia streams.

TCP-Friendliness of AWP. Let T_{tcp} , T_{awp} and T_{tfrc} denote the average throughput of AWP, TFRC and TCP connections respectively. The friendliness is defined as [8]:

$$F_{awp} = T_{awp}/T_{tcp} \quad F_{tfrc} = T_{tfrc}/T_{tcp} \quad (14)$$

Note that the closer to 1 the result is, the more TCP-friendly the mechanism is. From Fig 7(b) it can be inferred that AWP is more friendly than TFRC on average, which is very important when network sources run different protocols.

5 Conclusion

In this paper we propose a novel multiple time scale congestion control scheme for multimedia traffic. It exploits self-similar property of network traffic and makes predictions of network available bandwidth to enhance congestion control efficiency. We have examined and evaluated AWP scheme carefully via ns simulation. The results show that AWP has achieved the following performance: (1) high network bandwidth utility with low packet loss event probability; (2) good adaptation to the variation of network environment; and (3) smoothness and TCP-friendliness for transfers. All these features are very important for the QoS improvement of multimedia transmission and stability of the network.

References

1. J. Widmer, R. Denda, and M. Mauve, "A Survey on TCP-friendly Congestion Control", *IEEE Network Magazine*, Vol.15, No.3, May, 2001.
2. T. Tuan, and K. Park, "Multiple Time Scale Congestion Control for Self-similar Network Traffic", *Performance Evaluation* 36-37(1-4), pp. 359-386, 1999.
3. V. Riberiro, M. Coates, R. Riedi, S. Sarvotham, et al, "Multifractal Cross-Traffic Estimation" *Proceeding ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, September 2000.
4. S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation-Based Congestion Control for Unicast Applications", *Proceedings of ACM SIGCOMM*, 2000.
5. D. Bansal and H. Balakrishnan, "Binomial Congestion Control Algorithms" *Proceeding of IEEE INFOCOM*, April 2001.
6. D. Bansal and H. Balakrishnan, "TCP-friendly Congestion Control for Real-time Streaming Applications", *Proceeding of IEEE INFOCOM*, April 2000.
7. R. Riedi, M. Crouse, V. Ribeiro and R Baraniuk, "A Multifractal Wavelet Model with Application to Network Traffic", *IEEE Transactions on Information Theory*, Vol 45, No. 3, April 1999.
8. Q Zhang, WW. Zhu, YQ. Zhang, "Network-adaptive Rate Control and Unequal Loss Protection with TCP-Friendly Protocol for Scalable Video over Internet", in *special issue selected from IEEE ICME'00 on Multimedia Communications Journal of VLSI Signal Processing - Systems for Signal, Image and Video Technology*, 2001.
9. S. McCanne and S. Floyd. Network Simulator-2. <http://www.isi.edu/nsnam/ns/>
10. E. Weisstein. MathWorld, <http://mathworld.wolfram.com>
11. P. Abry, P. Flandrin, M.S. Taqqu, and D. Veitch, "Wavelets for the Analysis, Estimation, and Synthesis of Scaling Data", in *Self Similar Network Traffic Analysis and Performance Evaluation*, K. Park and W. Willinger, Eds., Wiley, 2000.

Dynamic Programming Based Adaptation of Multimedia Contents in UMA

Truong Cong Thang, Yong Ju Jung, and Yong Man Ro

Multimedia Group, Information and Communication University (ICU)
119, Munjiro, Yuseong, Daejeon, 305-714, Korea
{tcthang, yjjung, yro}@icu.ac.kr

Abstract. Content adaptation is an effective solution to support the quality of service for multimedia services over heterogeneous networks. This paper deals with the accuracy and the real-time processing, two important issues in making decision on content adaptation. We present the content adaptation as a constrained optimization problem, considering both modality conversion and content scaling. To this problem, we propose an optimal algorithm and a fast approximation based on the Viterbi algorithm of dynamic programming. Through experiments, we show that the proposed algorithms can enable accurate adaptation decisions and can support the real-time requirement.

1 Introduction

In a universal multimedia access (UMA) system, content adaptation is an important method to provide the best possible presentation under various constraints of terminals and networks. Basically, content adaptation includes three major modules: decision engine, modality converter, and content scaler (Fig. 1). The decision engine analyzes the content description, user preferences, resource constraints and then makes optimal decision on *modality conversion* and *content scaling*. The modality converter and the content scaler include the specific converting and scaling operations to adapt (or transcode), either offline or online, the content objects according to instructions from the decision engine.

The functionality of the decision engine is essentially the same for both online and offline transcodings. It is important that the decision engine can both provide accurate decisions and support the real-time processing. However, these two related issues of decision engine have been addressed just little so far. In [1], several searching algorithms were proposed to find the appropriate adapted content version. In [2], the adaptation process was modeled as the resource allocation and then solved by the Lagrangian method. This method is supposed to have low complexity, however the accuracy cannot be guaranteed when the content value (utility) functions are non-concave. Also, various adaptation frameworks were proposed in [3][4]. These approaches are said to be effective, yet no processing time has been reported.

In our previous work [5][6], we have proposed an approach that accurately adapts both the *quality* and the *modality* of contents under different constraints

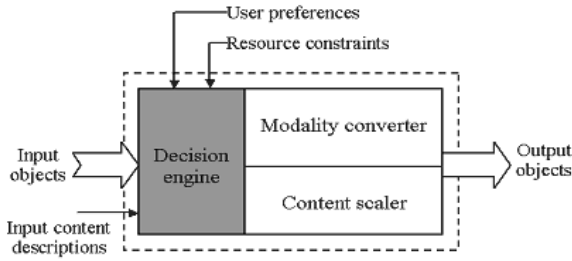


Fig. 1. Architecture of an adaptation engine

of terminal/networks and user preferences. In this paper, we show that the Viterbi algorithm of dynamic programming and a fast approximation can be used in this framework to find the optimal solution, and especially they are potential to support the real-time requirement of decision engine.

The paper is organized as follows. Section 2 reviews the problem formulation of content adaptation. In section 3, we propose the Viterbi-based algorithms to accurately solve the problem. Section 4 presents the experiment results on the performance of proposed algorithms. Finally section 5 concludes the paper.

2 Problem Formulation

Let us first define some basic terms used in this paper. A *multimedia document* is a container of multiple *content objects*. A content object (or *object* for short) is an entity conveying some information, e.g. a football match. An object may have many *content versions* of different modalities and qualities. A content version is an instance of the object, e.g. a video file showing a football match.

To adapt a document to some constraint, the decision engine has to answer two basic questions for every object: 1) which is the modality of output object and 2) what is the quality (content value) of output object. Without answers to these questions, we cannot apply the appropriate operations of modality conversion and content scaling.

The decision-making process of the decision engine can be represented as the constrained optimization problem as follows [2][5][6]. Suppose we have a document consisting of multiple objects. Let denote R_i and V_i the resource and content value of the object i in the document. The content value V_i can be represented as a function of resource R_i , modality capability M of client terminal, and user preference P_i :

$$V_i = f_i(R_i, M, P_i). \tag{1}$$

Then the problem of content adaptation for the given document is as follows:

Given a resource constraint R^c , find the set of $\{R_i\}$ so as

$$\sum_i V_i \text{ is maximum, while } \sum_i R_i \leq R^c. \tag{2}$$

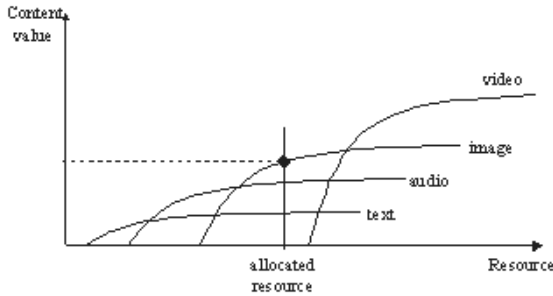


Fig. 2. Overlapped content value model of a content

Regarding equation (1), we have proposed the *Overlapped Content Value* (OCV) model to represent the relationship between content value, modalities, and resource [5][6]. Each object will be given an OCV model (Fig. 2) representing the content values of different modalities versus the resource. The number of curves in the model is the number of modalities the object may have. Here, the content value curve of each modality (called *modality curve*) can be assigned manually or automatically. Each point on a modality curve corresponds to a content version of that modality. The intersection points between the curves are the conversion boundaries between modalities. The final content value function is the upper hull of the model. A point on the hull is called a *selection*.

Denote K_i as the number of modalities and VM_{ij} as the content value curve of modality j of the content object i , $j=1\dots K_i$. The content value of an object, generally represented by (1), can be mathematically rewritten as follows:

$$V_i = \max\{w_{ij} \times VM_{ij} | j = 1 \dots K_i\} \tag{3}$$

where w_{ij} is the scale factor of modality j of object i and $j=1$ is the original modality.

Essentially, a modality curve represents the content scaling process of the object in the given modality [7], and the OCV model shows the relationship between modalities. An example of building the OCV model for the case of networked video was presented in [7]. Given a content value function for each object, a resource allocation method is then used to distribute the resource among multiple contents. Mapping the allocated resources back to content value models, we can find the appropriate qualities and modalities of adapted contents.

The use of OCV model to integrate the user preference P_i (on modality conversion) and the modality capability M into content value V_i was presented in [5][6]. This paper focuses on the optimization algorithms to make accurate decisions on modality conversion and content scaling. From now on, (3) is used instead of (1) to represent the content value. In fact, this constrained optimization problem is often solved by two basic methods: Lagrangian method and dynamic programming method. The Lagrangian method cannot provide accurate results when content value functions are non-concave [8]. Meanwhile, dynamic

programming can work with the non-concave functions to find exactly the needed version. An illustrative example on the accuracy of these two methods can be found in [8]. The disadvantage of dynamic programming is the high complexity. Yet, through practical considerations, we will show that the Viterbi algorithm is suitable for real-time operation of decision engine.

3 Solutions Based on Dynamic Programming

3.1 Optimal Solution by Viterbi Algorithm

A content value function can be continuous or discrete. If the function is continuous, we may discretize it because the practical transcoding is done in the unit of bits or bytes. Now we implicitly suppose the function is discrete. Then a content value function will have a finite number of selections. Meanwhile, function (3) is inherently non-concave, thus the above problem can be solved optimally by Viterbi algorithm of dynamic programming [8][9]. The principle of Viterbi algorithm lies in building a trellis to represent all viable allocations at each instant, given all the predefined constraints. The basic terms of Viterbi algorithm are defined as follows (Fig. 3):

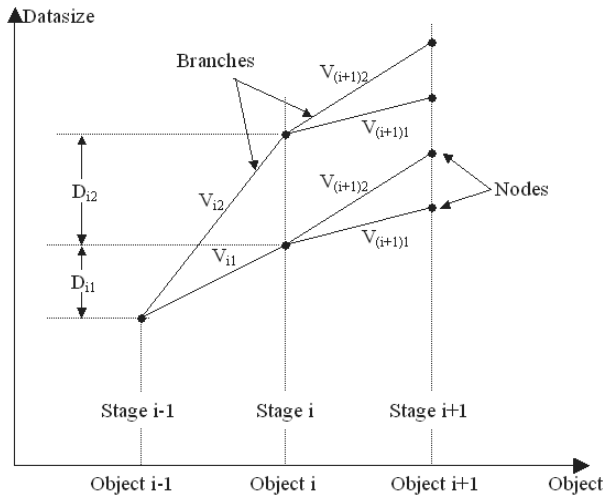


Fig. 3. Trellis diagram grown by Viterbi algorithm

- *Trellis*: The trellis is made of all surviving paths that link the initial node to the nodes in final stage.
- *Stage*: Each stage corresponds to an object to be adapted.
- *Node*: In our problem, each node is represented by a pair (i, a_i) , where $i = 0 \dots N$ is stage number, a_i is the accumulated resource of all objects until this stage.

- *Branch*: If selection k at stage i has the value-resource pair (V_{ik}, R_{ik}) , then node $(i - 1, a_i - 1)$ will be linked by a branch of value V_{ik} to node (i, a_i) with:

$$a_i = a_{i-1} + R_{ik}, \tag{4}$$

satisfying (if not, the branch will not be linked):

$$a_i \leq R^c. \tag{5}$$

- *Path*: A path is a concatenation of branches. A path from the first stage to the final stage corresponds to a set of possible selections for all objects.

From the above, we can immediately see that the optimal path, corresponding to the optimal set of selections, is the one having the highest accumulated content value. We now apply Viterbi algorithm to generate the trellis and then to find the optimal path as follows [9][10]:

Optimal algorithm:

- Step 0:** Start from the initial node $(0,0)$
- Step 1:** At each stage i , add possible branches to the end nodes of the surviving paths. At each node, a branch is grown for each of the available selections; the branch must satisfy condition (5).
- Step 2:** Of all the paths arriving at a node in stage $i+1$, the one having the highest accumulated content value is chosen, and the rest are pruned.
- Step 3:** Increase i and go to step 1.
- Step 4:** At the final stage, compare all surviving paths then select the path having the highest accumulated content value. That path corresponds to the optimal set of selections for all objects.

3.2 Fast Approximation Algorithm

Obviously, the complexity of above algorithm decreases if the number of selections is reduced. For this purpose, one can intuitively omit or merge some neighboring selections. We note that in the searching process at each stage, if the current selection has a negligible content value change compared to the last un-omitted selection (called the last considered selection), this selection can be omitted. So, if we set a minimum threshold for the changes of the content values, we can reduce the number of selections considered for an object by omitting the selections having the content value changes within the threshold. Denote k^* the last considered selection, we modify the above step 1 to obtain a fast approximation algorithm as follows:

- Step 1:** At each stage i , add possible branches to the end nodes of surviving paths. At each node, do the following:
 - Step 1.1: Add branch for selection 0
 - Step 1.2: Check the selection k ($k > 0$): if $|V_{(i+1)k} - V_{(i+1)k^*}| > threshold$ and if (5) is satisfied, add branch for selection k and let $k^* = k$.
 - Step 1.3: Increase k and go to step 1.2

Meanwhile, other steps are unchanged. Actually, the modified step 1 constantly checks the content value changes and tries to connect a branch only if the content value difference between the current selection and the last considered one is greater than the threshold. This technique is applicable to any ranges of resource because the omittable selections can exist anywhere, especially at the saturate intervals.

4 Experiments

We develop a UMA test-bed in which the adaptation engine is built on a Windows2000 server with Pentium IV 1.7GHz and 256MB RAM. In our system, the content transcoding is done offline. Scaling operations include reducing the spatial size for video and image modalities, reducing bandwidth for audio, and truncating the words for text. The resource constraint is datasize constraint D^c and the objects are transcoded in unit of kilobytes (KBs). For experiments, we employ a document consisting of six objects: one video, one audio, three images, and one text paragraph. The original datasizes of the objects are respectively 1500KBs, 480KBs, 731KBs, 834KBs, 813KBs, and 8KBs. The modality curve j of object i is modeled by the following analytical function:

$$VM_{ij}(R_i) = x_{ij}(R_i - y_{ij}) / (R_i - y_{ij} + z_{ij}) \quad (6)$$

where y_{ij} is the starting point, z_{ij} controls the slope, and x_{ij} is the saturate value of the function. The content values of the six objects are respectively 10, 3.0, 5.5, 6.0, 6.5, and 1.5, assigned according to their relative importance.

To examine the response of the adaptation, we vary the constraint D^c . The results are shown in Table I. In this table, the first column is D^c ; each object has two columns, one for the datasize and the other for the modality; the last column is the total content value of adapted document. Here, V, I, A, T mean video, image, audio, text modalities respectively. We can see that as D^c decreases, the datasizes of the objects are reduced to satisfy the datasize constraint of the whole document. Also, at some points, the modalities of the objects are converted to meet the constraint and to give the highest possible total content value.

Now we check the performance, including the processing time and the optimality, of the decision engine. First we employ optimal algorithm. The continuous lines in Fig. 4(a) and Fig. 4(b) show respectively the total content value of the adapted document and the processing time (to find the optimal solution) versus the different values of D^c . We see that, using optimal algorithm (threshold=0), the processing time is smaller than 0.5s, which is acceptable to the real-time requirement. This is actually because there are only six objects in the document.

Next, to reduce the processing time, we try to apply the fast algorithm with step 1 modified. The ‘‘good thresholds’’ are estimated by considering the content value and the processing time versus different thresholds. These relationships are depicted in Fig. 5 for three values of D^c , namely 4500KBs, 2500KBs, and 700KBs. From this figure we can guess that, when the threshold is around 0.02, the complexity may be reduced to one third while the total content value remains

Table 1. Results of the adapted documents with different values of constraint

D^c (KBs)	Object 1		Object 2		Object 3		Object 4		Object 5		Object 6		Content value
	D_1 (KBs)	Mod	D_2 (KBs)	Mod	D_3 (KBs)	Mod	D_4 (KBs)	Mod	D_5 (KBs)	Mod	D_6 (KBs)	Mod	
3000	1041	V	260	A	544	I	571	I	576	I	8	T	28.30
1000	343	V	90	A	179	I	189	I	191	I	8	T	24.08
300	100	V	30	A	52	I	56	I	57	I	5	T	14.88
200	85	V	26	A	20	A	47	I	21	A	1	T	11.38
130	45	I	25	A	19	A	20	A	20	A	1	T	8.47
110	36	I	21	A	17	A	17	A	18	A	1	T	7.59
90	34	I	5	T	16	A	17	A	17	A	1	T	6.61
70	14	A	5	T	16	A	17	A	17	A	1	T	5.61
10	1	T	3	T	1	T	2	T	2	T	1	T	1.27

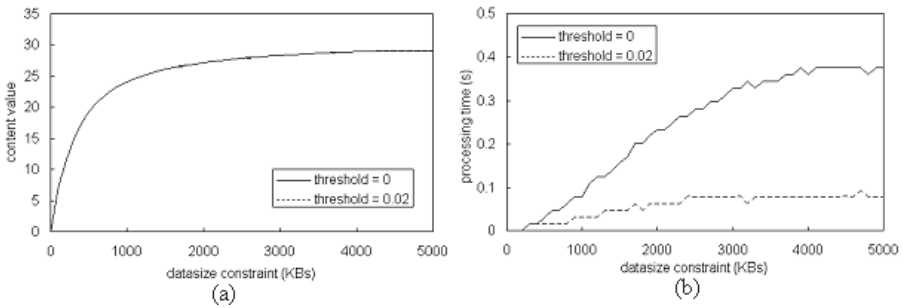


Fig. 4. Performance comparison of the decision engine using optimal & fast algorithms

nearly the same. The dashed lines in Fig. 4(a) and Fig. 4(b) show the performance of the decision engine using fast algorithm with threshold = 0.02. Now we can see that the processing time is much reduced, i.e. below 0.1s compared to 0.4s of optimal algorithm; meanwhile, total content value decreases very little. Besides, the thresholds seem to have more effect when D^c is high. This is because with the small values of D^c (e.g. $D^c=700$ KBs), the datasizes of adapted objects lie in small-value range, where the slope of content value function is often high, or the content value differences of adjacent selections are often larger than the thresholds.

In above experiments, there are only six objects in the document. Now we add some more image objects (which are the most popular contents on the web) to see how the processing time depends on the number of objects. The datasize of each added image objects is 900KBs. The datasize constraint is now set to be rather high, $D^c=5000$ KBs. Fig. 6 shows the relationship of the processing time versus the number of objects for three cases: threshold = 0 (optimal algorithm), threshold = 0.02, and threshold = 0.09 (fast algorithm). We see that when the number of objects is more than 20, the processing time of the original algorithm is more than 3s, meanwhile the result of the fast algorithm is very interesting. With threshold=0.09, the processing time for the document of as many as 30

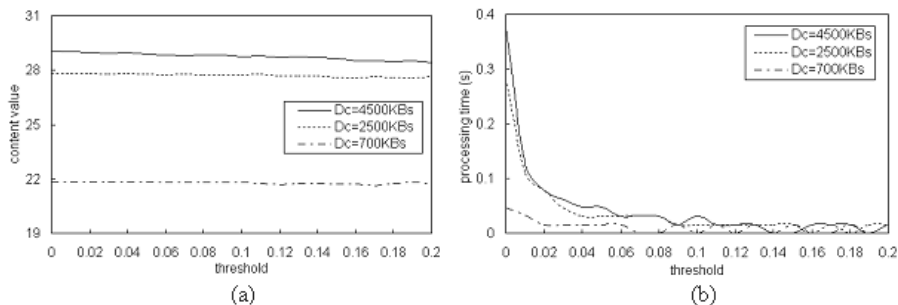


Fig. 5. Content value (a) and processing time (b) vs. threshold

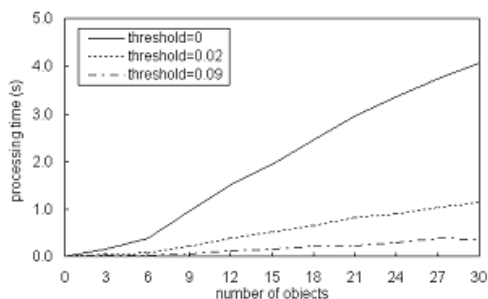


Fig. 6. Processing time vs. number of objects

objects is still below 0.5s. It should be noted that normally the acceptable delay for web browsing can be up to 11s.

From the experiments, we see that the computational complexity of the decision engine depends on three factors: the number of “considered selections” for each object, the value of the resource constraint, and the number of objects. In fact, the number of selections for an object is generally not many, about several thousands or several hundreds (even less), which is quite acceptable for fast searching. The threshold technique can reduce the number of considered selections by omitting the “negligible” ones. Furthermore, as in the discussion with Fig. 5, this technique is especially effective for the cases when the value of the resource constraint is high. In practice, the thresholds can be empirically estimated and stored as parameters of content description. As for the number of objects, a special point is that the number of objects to be transcoded in a multimedia document is not many, normally not more than several dozens. This is the main reason for the real time support of the decision engine. Moreover, the less important objects can be removed to reduce the number of transcoded objects. Taking advantage of these, the Viterbi algorithm can be well applied to decision engine for both real-time processing and accurate results.

5 Conclusions

We have presented a systematic approach, based on the overlapped content value model and dynamic programming, to tackle the content adaptation problem. The mechanism of the decision engine was formulated as a constrained optimization problem. Then Viterbi algorithm of dynamic programming and a fast approximation were employed to solve this problem optimally (and sub-optimally). The algorithms were shown to be effective for accurately adapting multimedia contents to different constraint values while meeting the real-time requirement. In the future, the solution will be extended to support multiple resource constraints. Also, the estimation of content value across different modalities is being studied.

Acknowledgements. This research was supported in part by the Digital Media Lab.

References

1. Lum, W.Y., and Lau, F.C.M.: A QoS-sensitive Content Adaptation System for Mobile Computing, Computer Soft. and Appli. Conference, (2002) 680-685
2. Mohan, R., Smith, J.R., Li, C.-S.: Adapting Multimedia Internet Content for Universal Access, IEEE Trans. Multimedia, Vol. 1, No. 1, (1999) 104-114
3. Chen, J., Yang, Y., Zhang, H.: An Adaptive Web Content Delivery System, Int. Conf. on Adaptive Hypermedia and Adaptive Web-based Systems, (2000) 284-288
4. Mérida, D., Fabregat, R., Marzo, J.L.: SHAAD: Adaptable, Adaptive and Dynamic Hypermedia System for Content Delivery, Workshop on Adaptive Systems for Web-based Education, Malaga, (2002)
5. Thang, T.C., et al.: CE Report on Modality Conversion Preference Part-I, ISO/IEC JTC1/SC29/WG11 M9495, Pattaya, (2003)
6. Thang, T.C., Jung, Y.J., Ro, Y.M.: Modality Conversion in Content Adaptation for Universal Multimedia Access, Int. Conf. on Imaging Sci., Syst. and Tech., (2003) 434-440
7. Thang, T.C., Jung, Y.J., Lee, J.W., Ro, Y.M.: Modality Conversion for Universal Multimedia Services, Workshop on Image Analysis for Multimedia Interactive Services, Lisboa, (2004)
8. Ortega, A., and Ramchandran, K.: Rate-distortion Methods for Image and Video Compression, IEEE Signal Processing Magazine, (1998) 23-50
9. Ortega, A., Ramchandran, K., Vetterli, M.: Optimal Trellis-Based Buffered Compression and Fast Approximations, IEEE Trans. Image Processing, vol. 3, (1994) 26-40
10. Forney, G.D.: The Viterbi Algorithm, Proc. IEEE, vol. 61, (1973) 268-278

Performance Analysis of MAC Protocol for EPON Using OPNET

Min-Suk Jung¹, Jong-hoon Eom², and Sung-Ho Kim¹

¹ Dept. of Computer Engineering Kyungpook National University,
Daegu, Korea

msjung@mmlab.knu.ac.kr, shkim@bh.knu.ac.kr

² Telecommunication Network Laboratory, KT, Daejeon, Korea
jheom@kt.co.kr

Abstract. An Ethernet PON (Passive Optical Network) is an economical and efficient access network that has received significant research attention in recent years. A MAC (Media Access Control) protocol of the PON, the next generation access network, is based primarily on TDMA (Time Division Multiple Access). In this paper, we addressed the problem of a dynamic bandwidth allocation in QoS (Quality-of-Service) based Ethernet PONs. We augmented the bandwidth allocation algorithms to support QoS in a differentiated service framework. Our differentiated bandwidth guarantee allocation (DBGA) algorithm allocates effectively and fairly the bandwidths between end-users. Moreover, we showed that a DBGA algorithm that perform weighted bandwidth allocation for high priority packets results in a better performance in terms of average and maximum packet delay, as well as the network throughput compared with some other dynamic allocation algorithms. We used the simulation to study the performance and validate the effectiveness of the proposed algorithm.

1 Introduction

In the last decade, the bandwidth in the backbone networks has significantly increased up to several Tbps by using WDM technology. The bandwidth in the local area networks also has increased up to 10Gbps based on Ethernet. Though access networks have improved their bandwidth using xDSL, Cable modem, etc., their bandwidth has been limited to several Mbps. Their limited bandwidth causes the bottleneck phenomenon in the access networks. Many service providers have already attempted to construct high speed access networks with the ultimate goal of FTTH [1][2][3]. The xDSL is currently used as a high-speed subscription access network. Although this type of network model is effective for an area where the demand is high, it is ineffective when the demand is only sporadic, plus its transmission capacity is inadequate for the latest multimedia services. Accordingly, a PON is a new technology for subscription access networks that supports the ability to connect subscribers using FTTx over the distances of up to 20km. In addition, a PON has the capacity to simultaneously

connect the maximum number of 64 ONUs through an optical splitter using an optical cable. Consequently, a PON has the ability to support the transmission of large amounts of data and is a very economical way to construct networks [4][5][6]. Among the various methods of transmitting data based on a PON, the first developed method was an ATM PON. Yet, when constructing a new network, using Ethernet has two advantages: Competitive price and home networks are constructed using the Ethernet. As such, the IEEE 802.3 EFM SG (Ethernet in the First Mile Study Group) was established with the goal of standardizing EPONs.

In this paper, we propose a method that controls the ratio of high-priority, medium-priority, and low priority without the additional bandwidth for high-priority traffic to each ONU, so as to minimize the delay of high-priority traffic. In addition, we design a simulation model whereby our proposed method is applied to the Ethernet PON and verify our proposed method is verified by analyzing the end-to-end delay and the throughput.

The rest of the paper is organized as follow. Section 2 describes the fashion of the study in Ethernet PON. In Section 3, we propose an efficient dynamic bandwidth allocation method. We will simulate our model and analyze its results in Section 4. Section 5 offers the conclusion and describes future research.

2 Ethernet PON

In this section, we describe differences between the Ethernet PON and ATM PON. We also explain the fundamentals of the operation in Ethernet PON. Figure 1 shows the EPON system structure, as suggested by the IEEE 802.3 EFM SG. The OLT and the ONU are located at the End Point of a PSS (Passive Star Splitter), whereby they are connected by an optical fiber. The PON is either distributed into several identical optical signals or is united into one signal according to the transfer direction of the optical signal. PSSs are economical as they are easy to construct, require little maintenance and are inexpensive to repair cost. Also, since a PSS is a passive component, it does not require any extra power supply. In addition, since the OLT and the ONU are connected by a Point-to-Multipoint form, the installment cost of the optical fiber is lower than that of a Point-to-Point form.

To compare the Ethernet PON with an ATM PON, a very different point is the transmission data type is required. Ethernet PON is able to transmit the variable length IP packets but the ATM PON only transmits the fixed length (53bytes) cells [8]. Moreover, the Ethernet PON is able to expand easily the channel speed up to 10 Gbps, but the ATM PON is limited in 622Mbps. Ethernet PON broadcasts traffic in the downstream direction, because the Ethernet PON is Point-to-Multipoint in the down-stream, but it is Multipoint-to-Point in the upstream. Thus, it requires a special MAC protocol to avoid an upstream traffic collision. To share the upstream link, Ethernet PON uses TDMA MAC protocol.

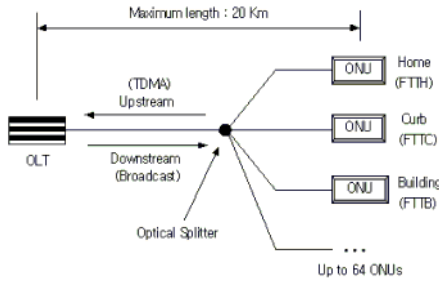


Fig. 1. The structure of the Ethernet PON is proposed in IEEE 802.ah

3 The Bandwidth Allocation Algorithm

In Figure 2, we present the bandwidth allocation procedure. The bandwidth allocation procedure is as follows: First, an ONU notifies its buffer status using a REPORT message. After the OLT receives a REPORT message, it executes the bandwidth allocation algorithm and grants their allocated bandwidth using GATE messages to each ONU. The bandwidth management plays an important role in guaranteeing QoS. The proposed bandwidth allocation algorithms in [7] are fixed, limited, credit, and gated. Here, they applied a strict priority scheduling mechanism (defined 802.1D) to the limited algorithm which is estimated higher than the other bandwidth allocation algorithms in [7]. Low-priority class traffic, however is not able to guarantee a sufficient bandwidth at the high load. As a result, low-priority class traffic may suffer from a starvation. In [9] they propose a DBA2 algorithm that combines the predicted credit based method with a priority-based scheduling mechanism.

The method that adds the predicted credit to the request time for high-priority traffic may reduce the packet delay in this cycle, but the extended cycle time influences the next cycle time. Finally, it will accumulate so that it increases the queuing delay. In addition, if the predicted credit is not used, the throughput

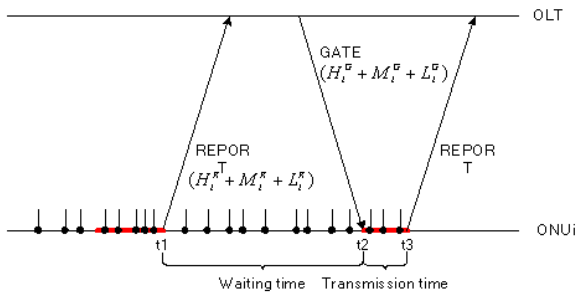


Fig. 2. The Bandwidth Allocation Procedure in Ethernet PON

will decrease. Therefore, we propose a DBGA (Differentiated Bandwidth Guarantee Allocation) algorithm that is fair and does not increase the queuing delay for high-priority class traffic.

B_i^R represents the requested bandwidth from ONU i and consists of high-priority, medium-priority, and low-priority traffic. It is as follows:

$$B_i^R = H_i^R + M_i^R + L_i^R \quad (1)$$

If we represent B_i^R as the granted bandwidth to ONU i , as below:

$$B_i^G = (H_i^R + L_i^R(1 - \omega)) + M_i^R + L_i^R\omega \quad \text{for} \quad W_i^{MIN} \leq B_i^G \leq W_i^{MAX} \quad (2)$$

ω represents the yield ratio whereby the low-priority class gives the high-priority class a part of itself. If the OLT allows the transmission of all packets in each ONU's buffer. Then, the ONU that has the most packets will monopolize the upstream bandwidth. To avoid this situation the OLT limits the maximum transmission window size W_i^{MAX} and the minimum transmission window size W_i^{MIN} . So, W_i^{MAX} guarantees the maximum transmission delay and W_i^{MIN} guarantees the minimum bandwidth, thereby it is able to satisfy SLA (Service Level Agreement). To reduce the control packet scheduling time, the OLT allocates the bandwidth by dividing the total number of ONUs into two parts using an interleaving method. The algorithm is the following:

```
DBGAlgorithm (H_R[], M_R[], L_R[], i)
{
  if (i == |N/2|)
    for(i=1; i<|N/2|; i++)
      {
        H_G[i] = H_R[i] + L_R[i](1-w);
        M_G[i] = M_R[i];
        L_G[i] = L_R[i] - L_R[i]*w;
      }
  if (i == N)
    for (i=|N/2|+1; i<N; i++)
      {
        H_G[i] = H_R[i] + L_R[i](1-w);
        M_G[i] = M_R[i];
        L_G[i] = L_R[i] - L_R[i]*w;
      }
}
```

The detailed control messages scheduling the problem are described in the next section.

4 Simulation and Analysis

In this section, we analyze the simulation results of the different bandwidth allocation algorithms presented in the previous section. Our simulation model is

developed using OPNET. We consider a PON structure with 16 ONUs connected to a tree topology. The maximum distance between each ONU and the splitter is 20km. The channel speed is considered to be 1Gbps and the maximum cycle time is 2ms. The practical traffic generation is an important role for an accurate simulation. Therefore we support the following traffic classes:

1. EF (Expedited Forwarding): This class has the highest priority and is a typical voice data. EF has a constant data size of 24-bytes. Including Ethernet and UDP/IP headers results in a 70-bytes Ethernet frame generated every 125 μ s. The inter-arrival time follows the Poisson distribution.
2. AF (Assured Forwarding): This traffic class has a lower priority than EF traffic and is generated by VOD (Video On Demand) or video conference. AF frame size is increased from 64-bytes to 1518-bytes and follows the uniform distribution. The data-centric traffic such as AF is characterized by self-similarity and LRD (Long Range Dependence). Considering these characteristics, the traffic model is implemented in ON and OFF periods. The ON periods generate packets but the OFF period do not (silent). Both the ON and OFF periods follow Pareto distribution, and the inter-arrival time follows the exponential distribution within the ON periods.
3. BE (Best Effort): This class has the lowest priority and is the fully data-centric traffic. This class includes http, ftp, e-mail, etc. The traffic has properties the same as AF (self-similarity and LRD).

We simulated and analyzed the results of fixed, limited, DBA2 and the proposed DBGA allocation method about the three kinds of the data traffic. The experiments consider the current Internet traffic. Since a real-time service like VOD on the web occupies 70% of the total traffic, we generated a 1:1:2 traffic rate on the EF traffic, AF traffic and BE traffic. With that situation, we wanted to determine how the proposed algorithm worked. In the implemented experiment model, we applied a strict priority scheduling mechanism to the fixed allocation algorithm and a priority-based scheduling mechanism to the limited, DBA2 and DBGA allocation method.

Figure 3 shows (a) the maximum delay and (b) the average delay which impact high-priority traffic on each allocation algorithm. As shown in (a) and (b) of Figure 3, high-priority traffic has the end-to-end delay in the order of DBGA, Limited, DBA2 and Fixed allocation methods in the load which is below 0.7. Though DBA2 and DBGA are based on the limited allocation method, DBGA is what better than DBA2, because DBA2 method is based on credit, and the ONU requests more data than what is queuing currently. So, the total cycle time becomes longer.

The DBGA allocation method linearly allocates additional data that are borrowed from BE to EF. Accordingly, it does not have much influence on the entire cycle time. Figures 4 (a) and (b) show the maximum delay and average delay in medium-priority traffic. Regarding the four allocation algorithm methods does

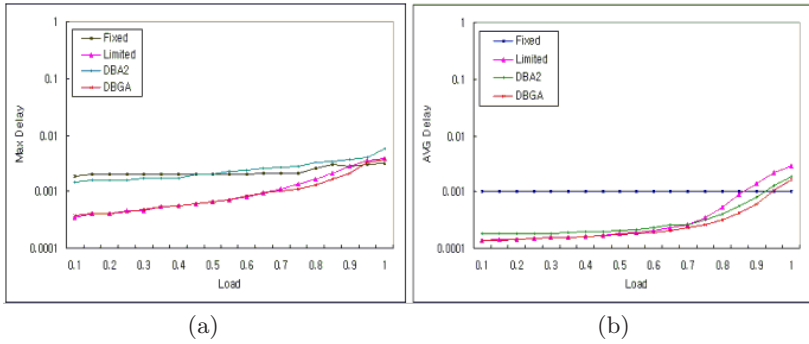


Fig. 3. The end-to-end delay of the high-priority class for each algorithm: (a) the maximum end-to-end delay and (b) the average end-to-end delay.

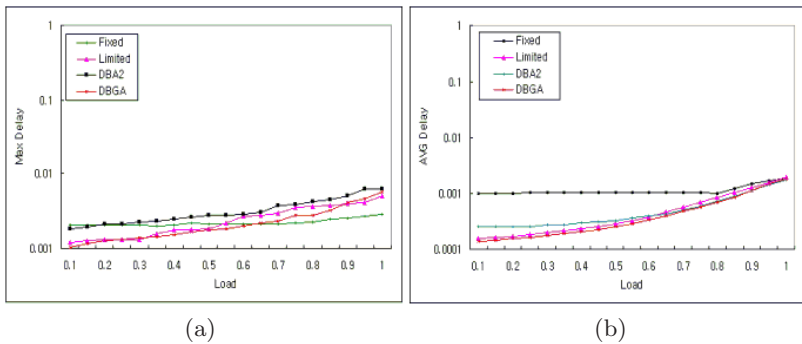


Fig. 4. The end-to-end delay of medium-priority traffic class for each algorithm: (a) the maximum end-to-end delay and (b) the average end-to-end delay.

not apply a special allocation algorithm on medium-priority traffic. Hence end-to-end delay is influenced by a cycle time.

Figures 5 (a) and (b) show the maximum delay and average delay of low priority traffic. The bandwidth allocation methods, except for the fixed bandwidth allocation method, show a similar end-to-end delay. In the fixed bandwidth allocation method, as the load increases, the queuing data and end-to-end delay increase faster than the others because of the unfairness of the strict priority scheduling mechanism.

Figure 6 represents the maximum and average end-to-end delays, when the DBGA is applied to the bandwidth allocation method. As described in Section 3, because the DBGA allocation method does not allocate the additional credit high-priority class, it gives high-priority class to a part of the real request of the low-priority class. It can prevent waste of bandwidth and keep a cycle time similar to the limited allocation method. It can guarantee the end-to-end delay of high priority traffic.

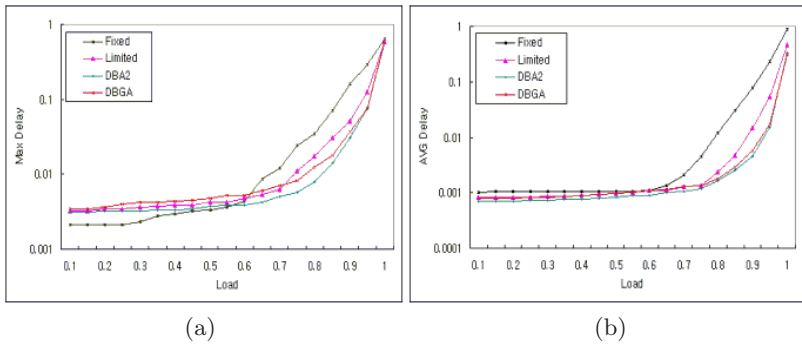


Fig. 5. The end-to-end delay of low-priority traffic class for each algorithm: (a) the maximum end-to-end delay and (b) the average end-to-end delay.

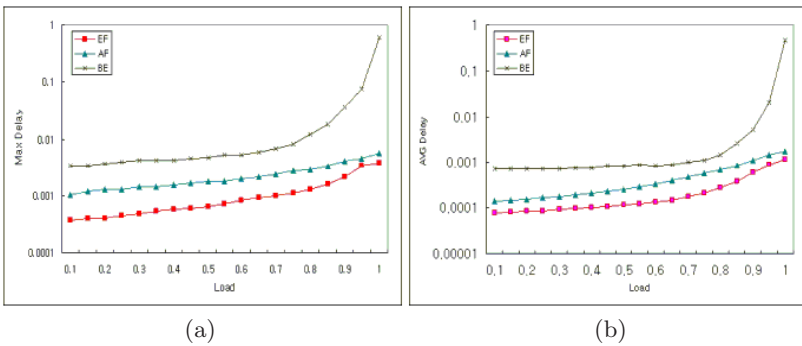


Fig. 6. The end-to-end delay of DBGA for each traffic class: (a) the maximum end-to-end delay and (b) the average end-to-end delay.

5 Conclusion

This paper dealt with the dynamic bandwidth allocation problem in the Ethernet PON. In particular, we proposed dynamic allocation algorithm supporting QoS in the differentiated service. The bandwidth allocation method is based on strict priority scheduling mechanism results in increasing delays for a specific traffic class. The method that allocates predicted credit to the high-priority service shows that it didn't improve the performance because of the increasing total cycle time. In order to correct this draw-back, we proposed a DBGA that adds a part of the timeslot for low-priority packets to the timeslot for high-priority packets.

When the proposed dynamic bandwidth allocation algorithm is compared with other algorithms, it shows that the proposed allocation algorithm is much greater than the average and maximum packet delay. Furthermore, the performance of the bandwidth allocation algorithm is dependant on the cycle time

as well as the timeslot size. So, the proposed algorithm that represented the shorter-cycle time in the appropriate range is a better performer than the long-cycle time. To verify the proposed algorithm, we used the OPNET simulation tool. Further study is required to verify about the performance in Ethernet PON, when the bandwidth is at 10Gbs. Also, we plan to analyze the performance in the DBGA algorithm, which is adapted in the system model using a control channel of the fixing transceiver in the WDM PON, and a data channel of the tunable transceiver.

References

1. A. Cook and J. Stern, "Optical fiber access Perspectives toward the 21st century," *IEEE Commun. Mag.*, pp. 78-86, Feb. 1994.
2. Y. Takigawa, S. Aoyagi, and E. Maekawa, "ATM based passive double star system offering B-ISDN, N-ISDN, and POTS," *Proc. of GLOBECOM'93*, pp. 14-18, Nov. 1993.
3. S. S. Kang, H. J. Kim, Y. Y. Chun, and M. S. Lee, "An Experimental Field Trial of PON Based Digital CATV Network," *IEICE Trans. Commun.*, Vol. E79-B, No. 7, pp. 904-908, July 1996.
4. G. Pesavento and M. Kelsey, "PONs for the Broadband Local loop," *Lightwave*, Vol. 16, No. 10, pp. 68-74, Sep. 1999.
5. B. Lung, "PON Architecture Future proof FTTH," *Lightwave*, Vol. 16, No. 10, pp. 104-107, Sep. 1999.
6. G. Kramer B. Mukherjee, and G. Pesavento, "Ethernet PON(Ethernet PON): Design and Analysis of an Optical Access Network," *Photo. Net. Commun.*, Vol. 3, No. 3, pp. 307-319, July 2001.
7. G. Kramer and B. Mukherjee, "IPACT: A Dynamic Protocol for an Ethernet Passive Optical Network (Ethernet PON)," *IEEE Commun., Mag.*, pp. 74-80, Feb. 2002.
8. ITU-T Recommendation G.983.1, "Broadband Optical Access Systems Based on PON," Geneva, Oct. 1998.
9. Chadi M. Assi, Y. Ye, Sudhir Dixit, Mohamed A. Ali, "Dynamic Bandwidth Allocation for Quality-of-Service Over Ethernet PONs," *IEEE Journal on Selected Areas in Communications*, 21(9):1467-1477, November 2003.

Adaptive FEC Control for Reliable High-Speed UDP-Based Media Transport

Young-Woo Kwon¹, Hyeyoung Chang², and JongWon Kim¹

¹ Networked Media Lab., Department of Information and Communications, Gwangju Institute of Science and Technology (GIST), Gwangju, 500-712, Korea
{ywkwon, jongwon}@gist.ac.kr

² Digital Media R&D, Samsung Electronics Co., Ltd, Suwon, Korea
hye.chang@samsung.com

Abstract. In this paper, we propose a reliable high-speed UDP-based media transport with an adaptive FEC (forward error correction) error control. The proposed adaptive transport scheme controls the amount of redundancy by monitoring the network so that it can effectively adapt to the fluctuations of underlying networks. The monitored feedbacks of the receiver enable the sender to aware of current reception status (i.e., rate/type of packet loss and delay change) and to estimate the expected network status. Based on this, the proposed media transport attempts to enable reliability by adaptively controlling the amount of both total sending rate and the FEC code ratio. Experiments with high-speed network have been conducted to verify the performance of the proposed transport that demonstrates the enhanced reliability at the speed of up to several hundred Mbps.

1 Introduction

To provide a reliable transport service for massive continuous media over high-speed networks, packet-level FEC (forward error correction) has been studied to overcome the randomized packet losses, i.e., packet erasures. Together with retransmission schemes that trade the bandwidth efficiency with additional delay, the FEC-based transport schemes are known to be very effective, especially when the latency is limited. However, when deployed in an open loop by simply fixing the redundancy level for given estimated loss-rate, the FEC-based schemes will end up with wastes in precious network bandwidth. It is thus important to react to the network fluctuations by adaptively controlling the amount of redundancy. To cope with packet losses as well as to avoid bandwidth waste, several adaptive FEC-based transports have been introduced to control the amount of redundancy adequately based on the monitored network feedbacks [1, 2].

However, most existing ideas for proactive packet-level FEC adaptation have been limited to low to medium speed ranges (e.g., up to several Mbps) and little attention has been paid to the issue of high-speed transport at the speed of up to Gbps. With the advent of optical-based high-speed networks, various efforts on the transport and application layers have been made over the

years to remove throughput limitation caused by the congestion control and others. Most recent researches have focused on TCP while only a few UDP-based high-speed transports such as SABUL (simple available bandwidth utilization library), UDT (UDP-based data transfer), and Tsunami are developed [5, 6]. SABUL and Tsunami are adopting rate-based controls by combining UDP transport and TCP control. These works, however, focus only on the retransmission for its reliability control. Thus, as in the case of interactive real-time media streaming where only small-size receiver-side buffering is allowed (i.e., delay is stringently constrained), they may not be the best solution in sustaining packet loss-rate below user-specified target. When there exist persistent packet losses, the proactive error control (i.e., FEC) will be a natural choice.

In this paper, targeting the reliable real-time delivery of immersive and massive continuous media, we propose an adaptive FEC-based UDP transport and corresponding controls with network monitoring. Based on the feedback of network monitoring (i.e., loss rate and loss burstiness), the proposed adaptive control guides redundancy of media streaming. It controls the amount of redundancy in face of network fluctuations while considering the limit on the sending rate (i.e., like the available bandwidth). More specifically, the proposed scheme includes 1) the estimation of network condition based on the receiver feedback of network monitoring, 2) the rate adaptation about the available bandwidth considering current network status estimate, and 3) the decision on the FEC code ratio to be applied. Focusing on the impact of high-speed aspects to the proposed FEC adaptation, the performance of proposed transport is evaluated over a high performance networking testbed by evaluating the performance of real-time delivery.

The outline of the paper is as follows. After discussing the details of the proposed transport control in Section 2, we discuss its performance evaluations over the networking testbed in Section 3. Finally, we wrap up with the conclusion in Section 4.

2 Proposed Adaptive FEC Transport Control

As of today, it is believed that high-speed transport of Gbps range is only possible under a high-performance networking environment where the QoS (quality of service) is provided to certain level [7, 8]. That is, certain level of guarantee is required to realize transport approaching Gbps range. Thus, under this kind of relatively guaranteed garden like environment, the proposed adaptive transport adjusts the redundancy of UDP packets based on the control packets. As depicted in Fig. 1, it is divided into three major components: the monitoring and estimation module for network condition, the rate adaptation module, and the adaptive FEC decision module. First, the network status is monitored to capture the packet loss rate, loss type, and others. For this, by calculating the standard deviation within a suitable adaptation window, we grasp the current network status. For the rate adaptation, the total sending rate is controlled to meet desired quality level while avoiding bulky loss and delay variation, which are

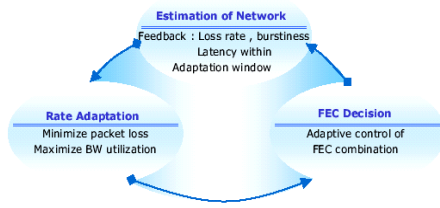


Fig. 1. Three major modules of the proposed adaptive transport.

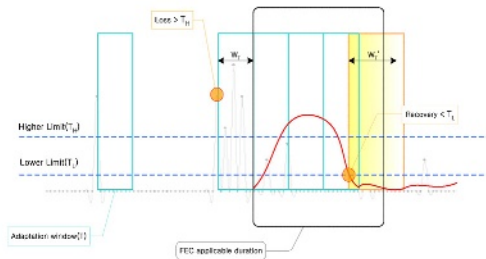


Fig. 2. Network estimation with the adaptation windows.

mainly linked with the estimated available bandwidth. The FEC decision module makes tradeoff between error resilience and bandwidth efficiency according to the fluctuation of current network. Special attention is also paid to mitigate the bandwidth overhead and the FEC encoding/decoding complexity that are undesirable for the high-speed transport.

2.1 Network Monitoring and Estimation

To capture the status of underlying networks, we monitor at the receiver such parameters as loss rate $L(t, T)$, and burst length $B(t, T)$. Here, t stands for sampling instant and T means the period of feedback. Based on regular feedbacks, the sender applies an adaptation window with duration W_T (set to multiples of T), as shown in Fig. 2, to estimate the network status smoothly. We also specify two fixed thresholds - *lower limit* (T_L) and *higher limit* (T_H) thresholds for packet losses. At first, when the loss begins to exceed T_H ¹, we enable the network estimation process at the sender.

Then, we calculate the standard deviation of loss $\sigma(L(t, W_T))$ and burst length $\sigma(B(t, W_T))$ within the adaptation window. Based on these variances,

¹ Currently, the sender doesn't apply the adaptation window while the current loss is maintained below T_H .

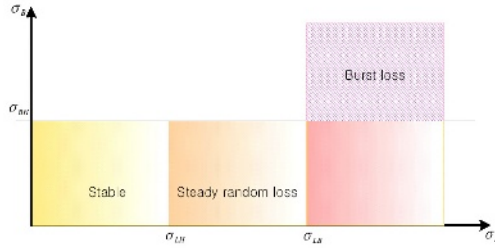


Fig. 3. Network states employed to differentiate adaptations.

we distinguish the network status into three cases - stable, steady random loss, and burst loss cases as depicted in Fig. 3. Note that σ_{LH} is the threshold for random loss standard deviation, σ_{LB} is the threshold for burst loss standard deviation, and σ_{BH} is the threshold for burst length. Also, to check the network variation further, we calculate the differentials of $L(t, T)$ like $f'_L(t_n) = (L(t_n + 1) - L(t_n)) / (t_{n+1} - t_n)$. When $f'_L(W_T)$ is above zero, loss rate is increasing and vice versa. Finally, let us comment more on the adjustment of adaptation window in Fig. 2. If some loss situation is maintained consistently in one window duration, the FEC is applied to combat these losses. With the application of FEC, recovery rate starts to increase while loss rate is decreasing. Similarly, the recovery rate is decreasing when the loss terminates. Thus, when the recovery rate meets T_L , we apply longer window $W'_T (\geq W_T)$ to allow time to converge. Note that, like this, the proposed adaptation window at sender helps us to check the persistence of network status change and avoid self-induced fluctuation resulting from oscillatory of FEC adaptation².

2.2 Rate Adaptation

In designing the high performance transport, it is also very important to consider the rate control principle. The transport has to adjust the sending rate by discovering the available bandwidth so that we can minimize the overall packet losses by avoiding the network congestion³. However, with the adaptive FEC, we are adding redundancy to the flow within the limit of available bandwidth. The consequence of this redundancy increase means that some portion of data should be omitted from the transport so that we can effectively reduce the required bandwidth. For the video case, increased redundancy can reduce residual packet losses at the cost of video quality. Especially with high-rate packet losses

² Note, however, that current version of network monitoring is still lacking in catching up with the rapid variations in the high-speed networks. In this work, we are only trying to explore the possibility of simplified monitoring to guide adaptation.

³ This issue is heavily related to the TCP-friendly congestion control. However, in this paper, we are barely touching this issue for the sake of simplicity.

with bursts, we should reduce the total sending rate while increasing the protection power at the same time.

Thus we adopt a rate-based gradual adjustment of sending rate to minimize the effect of fluctuation. That is, switching of rate does not happen instantly even though the distinct change of loss condition is predicted. The selected sending rate ($R(t)$), which is calculated by dividing the total amount of data with the whole time period spent, should be in the range of $B_{min} \sim B_{available}$. Here, $B_{available}$ is available bandwidth⁴ and B_{min} is a minimum guarantee bandwidth that is highly dependent on the type and bandwidth flexibility (i.e., adaptivity) of target real-time applications. In addition, the change of sending rate is related to the loss rate of underlying networks. Thus, measured loss rate after τ from time t can be expressed as a function of $R(t)$ as described in Eq. (1).

$$L(t + \tau) = f(R(t)), \quad B_{min} < R(t) < B_{available}. \quad (1)$$

The main role of function $f(R(t))$ is not only to limit loss rate under user specified level but also to maximize the bandwidth utilization. If there exists steady random losses, $R(t)$ is decreased reflecting the difference of current loss and the target limit T_H . On the other hand, when the losses are decreased, $R(t)$ recovers linearly back to $B_{available}$ in order to maximize the bandwidth utilization. However, if $R(t)$ gets close to the bandwidth limit, the loss rate may oscillate around T_H and can cause the unnecessary variations of sending rate. To solve this, $R(t)$ has to increase only up to $B_{available} - \Delta$. Also, to adapt to the network state in a gradual manner, we use $L(t, W_T)$ and $B(t, W_T)$. The augmented rate adaptation algorithm⁵ is given by Eq. (2).

$$f(R(t)) = \begin{cases} R(t) - \alpha(L(t, W_T) - T_H), & \text{if } L(t, W_T) > T_H \\ R(t) + \alpha, & \text{if } R(t) < B_{available} - \Delta \text{ and } L(t, W_T) < T_H. \end{cases} \quad (2)$$

2.3 FEC Decision

The proposed FEC decision scheme is summarized in Fig. 4. We start to send data at fixed (n_0, k_0) - (total number of packets in a block, number of source packet in a block) - initially. Then, based on the network status, different procedures are applied. If the loss is steady and controlled, current combination will be maintained. In case of random loss, the amount of redundancy should be able to cover the burst loss. That is, the number of redundant packets $h (= n - k)$ ⁶ has to be greater than $B(t, T)$ and we need to find k with given h . Note that now the decision on k is depending on the loss rate. For the burst loss case, h

⁴ Note that, according to the popular equation-based TCP-friendly congestion controls, the available bandwidth is usually determined with loss rate, RTT, and MTU size. However, $B_{available}$ is controlled according to the loss rate only in this paper.

⁵ This kind of aggressive rate adjustment should be improved further to provide friendliness to TCP traffics.

⁶ In general, we put the FEC combination (n, k) - (total number of packets in a block, number of source packet in a block).

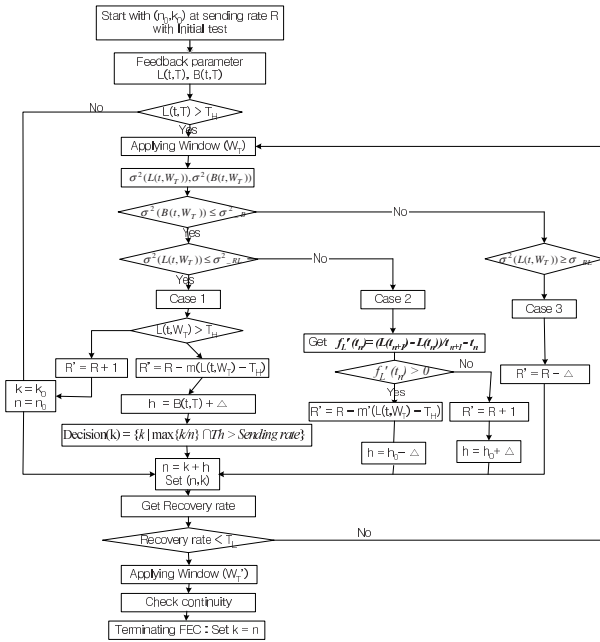


Fig. 4. Detailed procedure for FEC decision.

is increasing while k is decreasing to minimize the loss burstiness. However, the increase in k causes the reduction of protection ability as well as the increase of processing overhead. The protection efficiency for bandwidth is normalized by $k/n(\%)$. As an example, if one parity packet is generated over three packets (i.e., $(n, k) = (4, 3)$), the protection efficiency is said to be 75%. For the efficient utilization of bandwidth, the choice of FEC combination has to be determined to maximize the protection efficiency only under the condition that it satisfies user specified quality level. Therefore, when we can process the FEC faster than the required sending rate and the current packet loss is under T_H , we increase k linearly. In case of burst packet losses, where rapid decrease in total sending rate happens, h needs to be increased while k is decreasing. When the FEC combination is changed, the information (n, k) will be included in the FEC header so that the receiver can catch the updated FEC combination.

3 Experiments and Results

In Fig. 5, the building blocks of the proposed reliable high-speed transport system are depicted. At the sender side, source data are grouped into a block of k packets and varying number of parity packets, $h = (n - k)$, are attached by the FEC encoder. To enable high-speed transport, we exploit two independent threads for sending and feedback control. The receiver checks whether there is

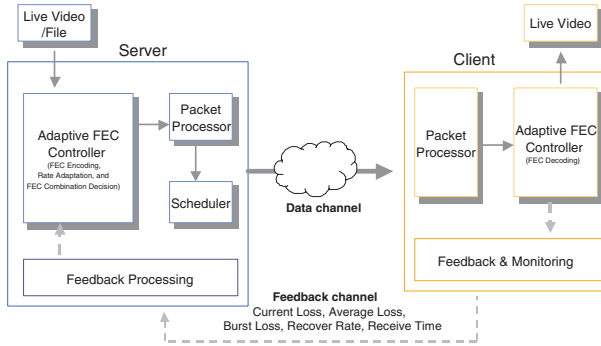


Fig. 5. The adaptive transport system.

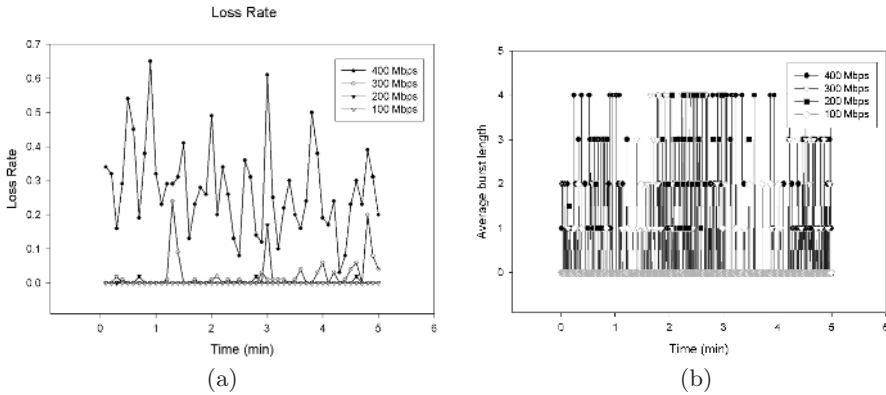


Fig. 6. (a) Loss rate (b) burst packet loss.

lost source packets by performing gap-based loss detection. Through this monitoring, the receiver sends back to the sender required feedback. If source packets are lost, wait until sufficient number (at least k) of packets arrive and then forwards them to the FEC decoder for reconstruction. Finally, packets are de-packetized, processed, and rendered to the display.

The performance of the proposed transport system is evaluated by transmitting over the real-world Internet. The WAN path of KOREN/KREONET between GIST (Gwangju) and KISTI (Daejeon) includes 3 hops in each direction and has an RTT of approximately 4 ms. Fig. 6 shows loss rate and burst length as a function of time that is measured between GIST and KISTI⁷. It is measured at daytime with the period of [0,5] min. Most of loss periods involve from one to five packets and it is confirmed by looking at the frequency distri-

⁷ At the time of this experiment, the link between GIST and KISTI is limited up to 430Mbps.

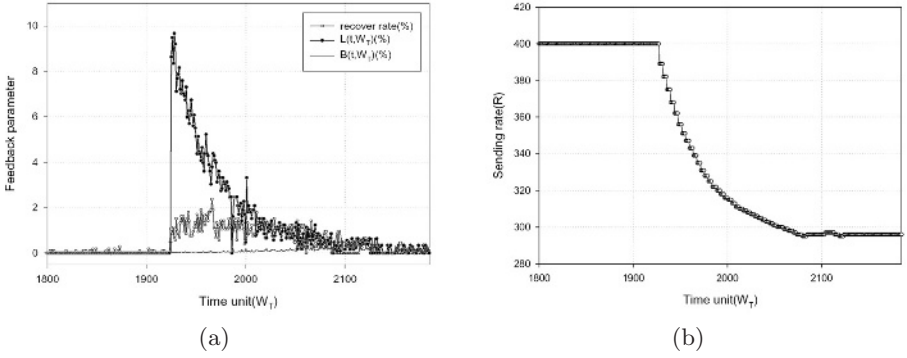


Fig. 7. (a) Loss rate (b) Sending rate.

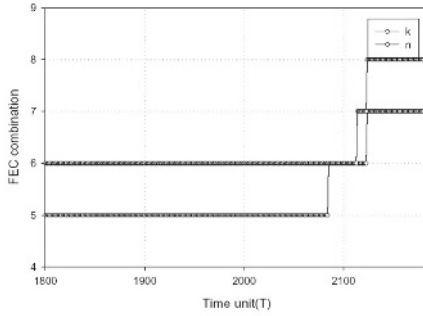


Fig. 8. Variation in the FEC combination.

bution in Fig. 6. We have done several experiments by increasing the sending rate over the same connection. The loss rate is increasing with the increase in the sending rate, and all cases have similar distribution as depicted above.

Fig. 7 shows the variation in the sending rate, and feedback parameter during time interval $[1800, 2200]$ time unit (W_T), which is interval of adaptation window. Depicted parameters are measured after classifying the network states. In this experiment, we set W_T as 1 sec, bandwidth offset Δ as 30Mbps, and reduction factor α is set to 1. To simulate network loads, we insert background traffic of 100Mbps at 1925 W_T . Initially, we start with the FEC combination (5,5) and assume T_H is set to 1%. Until 1925 W_T , the sending rate $R(t)$ is maintained at 400Mbps with small residual losses. Note that, to recover from these losses, h is set to 1. After inserting the background traffic, increased $L(t, W_T)$ starts to decrease if we reduce the total sending rate. The degree of reduction follows the difference amount of $L(t, W_T)$ and T_H . After FEC reconstruction, the recovery rate is also increasing. When the sending rate is saturated to about 295Mbps, $L(t, W_T)$ gets converged below 1% with less fluctuation. Like this, the

adaptive FEC is first applied when the loss is first detected as significant and the redundancy is controlled as depicted in Fig. 8. Over the time, the redundancy of FEC is adjusted automatically by reflecting the feedback parameters.

4 Conclusion

We have presented a reliable high-speed UDP-based transport and its adaptive FEC control. Based on the feedback, the adaptive FEC control can improve the quality of delay-constrained immersive media streaming while mitigating the impact of network fluctuation and system limitations over the high-speed networks. We are currently undergoing several refinements to address the limitations discussed in the current version of paper.

Acknowledgement. This work was supported by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting IT Research Center (RBRC) at Gwangju Institute of Science and Technology (GIST).

References

1. C. Padhyes, K. Christensen, and W. Moreno, "A new adaptive FEC loss control algorithm for voice over IP applications," in *Proc. IEEE International Performance, Computing and Communication Conference*, Feb. 2000.
2. K. French and M. Claypool, "Repair of streaming multimedia with adaptive forward error correction," in *Proc. SPIE Multicast Multimedia Systems and Applications (part of ITCOM)*, Aug. 2001.
3. D. Katabi, M. Hardley, and C. Rohrs, "Internet congestion control for future high bandwidth-delay product environments," in *Proc. ACM SIGCOMM*, Pittsburgh, PA, Aug. 2002.
4. S. Floyd, "High-speed TCP for large congestion windows," IETF RFC 3649, Dec. 2003.
5. Tsunami, <http://www.anml.iu.edu/anmlresearch.html>, retrieved on 09/01/2004.
6. H. Sivakumar, R. Grossman, M. Mazzucco, Y. pan, and Q. Zhang, "Simple available bandwidth utilization library for high-speed wide-area networks," *J. Supercomput.*, 2003.
7. A. Falk, T. Faber, J. Bannister, A. Chien, R. Grossman, and J. Leigh, "Transport protocols for high performance," *Communications of the ACM*, vol. 46, Nov. 2003.
8. I. Foster, M. Fidler, A. Roy, V. Sander, and L. Winkler, "End-to-end quality of service for high-end applications," *Computer Communications: Special Issues on Network Support for Grid Computing*, 2002.

Efficient Overlay Network for P2P Content Sharing Based on Network Identifier

Chanmo Park and JongWon Kim

Networked Media Lab. Department of Information and Communications,
Gwangju Institute of Science and Technology (GIST), Gwangju, 500-712, Korea
{cmpark, jongwon}@netmedia.gist.ac.kr

Abstract. In this paper, we propose an efficient overlay network for P2P (peer-to-peer) content sharing by incorporating network identifier of each peer in the construction of CAN (content addressable network)-variant DHT (distributed hash table)-based overlay. The network identifier that partially reflects the locality of each peer within the global Internet can help us to aggregate distributed peers towards more structured set of virtual peers. Through myns-based simulations coupled with GT-ITM topology, the improved efficiency of the proposed overlay construction is verified.

1 Introduction

P2P (peer-to-peer) concepts and networks for sharing files among peers have been popular since Napster [1] started its service at mid 1999. With the popularity of P2P, many researchers have been interested in media streaming systems based on P2P file sharing that serve as a directory server for locating the file and the owner of it. To discover contents for P2P media sharing systems, centralized or distributed approaches might be used. Napster, the centralized approach, has a centralized server that serves as a directory server and clients that download files by sending queries to nodes that own the file. In general, distributed content discovery approaches based on overlay network can be classified into two categories: unstructured and structured approach. Under unstructured Gnutella and KaZaA [1], all peers simultaneously take roles of a client, a server, and a message routing node on the overlay network. Structured approaches [1] such as Chord, Tapestry, Pastry, and CAN (content addressable network) [2] construct structured overlay network using DHT (distributed hash table). The structured overlay networks with DHT are based on the virtual coordinate space into which contents are mapped by hash functions. Thus, peers on the structured overlay network with DHT store content information (i.e., pointers to contents) on other nodes according to the assigned range within the virtual coordinate space. Routing messages are exchanged by forwarding along neighbor peers whose range are closer to a destination node. Therefore, the performance of structured overlay network with DHT is determined by the number of hops needed to route messages to the destination node. It is important to improve the performance of

content discovery to enhance P2P-based media streaming systems [3,4]. In this paper, we focus on improving the underlying P2P content sharing overlay network with DHT by reducing the number of hops for message routing. We use the network identifiers of participating peers to aggregate them into groups. In the proposed approach, the adapted virtual coordinate space is 2-dimensional and is occupied by participating peers according to the mapping rule based on their network identifiers. Participating peers having the same network identifier are mapped into the same virtual coordinate space. We expect this grouping to help as to reduce the number of hops for message routing.

The remaining part of this paper is organized as follows. In Section 2, we briefly review DHT's focusing on the CAN. The proposed overlay construction with the network identifier is described in Section 3. Section 4 shows the simulation environment and results. Finally, we conclude the paper in Section 5.

2 Background and Related Works

The CAN [2] is based on a d -dimensional Cartesian coordinate space on a d -torus. The Cartesian space of CAN is completely logical. The entire coordinate space is dynamically partitioned among all the nodes in the system such that every node owns its individual, distinct zone within the overall space. The virtual coordinate space is used to store (*key, value*) pairs as follows: to store a pair (K_1, V_1) , key K_1 is deterministically mapped onto a point P in the coordinate space using a uniform hash function. The corresponding (*key, value*) pair is then stored at the node that owns the zone within which the point P lies. To retrieve an entry corresponding to key K_1 , any node can apply the same deterministic hash function to map K_1 onto point P and then retrieve the corresponding value from the point P . If the point P is not owned by the requesting node or its immediate neighbors, the request must be routed through the CAN infrastructure until it reaches the node in whose zone P lies. Efficient routing is therefore a critical aspect of a CAN. Nodes in the CAN self-organize into an overlay network that represents this virtual coordinate space. A node learns and maintains the IP addresses of those nodes that hold coordinate zones adjoining its own zone. This set of immediate neighbors in the coordinate space serves as a coordinate routing table that enables routing between arbitrary points in this space. However, it is known that the resulting performance of CAN may suffer inefficiency problem when the number of zones is increased too much (as it adds a zone whenever a new node joins) [5]. To improve the routing efficiency of the CAN, one approach is to reduce each CAN-hop latency by using network proximity (i.e., via landmark-based or DNS-based measurement) [2]. The other is to reduce routing path length (i.e., number of hops) by increasing dimension of virtual coordinate space or by increasing reality (i.e., coverage) of each zone [5]. Although the former makes routing among zones to be effective in terms of the latency stretch (i.e., the ratio of CAN routing delay to IP routing delay), the construction of overlay network requires additional efforts to reflect the physical network topology. However, in this case, improvement in routing is

limited because of little knowledge about neighboring zones [6]. On the contrary, the latter can easily boost the performance of routing while keeping CAN's low maintenance cost. It also allows us to construct the overlay network without considering the physical network topology [5]. In this paper, we are mixing both approaches to improve routing performance. The proposed overlay network is reflecting network proximity by adopting network identifier. At the same time it propose an overlay network that consists of bigger size zones. Thus, the actual design needs to pursue both approaches with correct balance.

3 P2P Content Sharing Overlay Network Based on Network Identifier

3.1 Design Overview

The proposed overlay network is motivated by reducing the number of zones compared to CAN-based one. The proposed overlay network uses virtual 2-dimensional coordinate space with a distributed hash table to store $(key, value)$ pairs ranging from 0 to $2^{16} - 1$ in term of x and y coordinate, respectively. Nodes which are willing to connect a overlay network are shown in Fig. 1(a). To reduce the number of zones in CAN-based overlay, we aggregate nodes of similar characteristics into a zone. As a feature to control aggregation, we introduce network identifier of each peer which is composed of 4 bytes. Note that in general nodes in a LAN (local area network) shares the same network identifier. For that reason, we use network identifier as a factor to aggregate nodes into a zone. Thus, nodes with the same network identifier are aggregated into a zone. To allocate peers and contents on the 2-dimensional space, contents and peers are assigned ContentIDs and NodeIDs respectively. Contents are hashed up into ContentIDs (values ranging from 0 to $2^{32} - 1$) and then mapped into 2-dimensional space with ContentIDs. To map NodeIDs and ContentIDs into x and y coordinate (each 2 bytes), they are separated with higher and lower 2 bytes. Each 2 bytes are corresponding to x and y coordinate on the virtual 2-dimensional coordinate

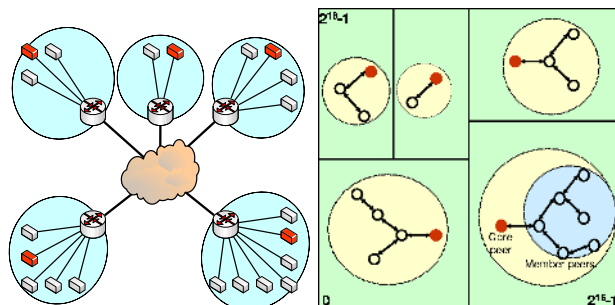


Fig. 1. (a) Network topology. (b) Proposed overlay network for P2P content sharing based on network identifier.

space. Peers are mapped into a zone, the virtual coordinate space, according to their NodeIDs (network identifiers) and own the zone. Although the proposed 2-dimensional space is originated from the concept of CAN, there are differences. Our approach uses fixed dimensional space, 2-dimension. But CAN uses n-dimensional space. We intentionally reduce the dimensional space to help peers' network identifiers to be mapped into space intuitively. To reduce the number of routing hops, mapping of peers in our approach provides aggregation of peers with the same network identifiers into a zone.

A zone owns its individual and distinct space within the overall space. A zone is composed of only one core peer and member peers because there are nodes with the same network identifier. If the number of peers is larger than one within a zone, peers store the same (*key, value*) pairs on their own storages and should maintain an overlay network between a core and member(s). A core peer represents the his zone to neighboring zones and exchanges messages between neighboring zones like a peer in CAN with hiding members in his zone. If there are member nodes in a zone, a shared binary tree is constructed and used to exchange messages within the zone. Peers in proposed overlay network are required to keep hash table to store content index over the content sharing overlay network, peer's zone range, core peer's information (IP address and port number), and a routing table which includes neighboring zones and member peers in his zone. Fig. 1(b) shows the proposed overlay network which is constructed with Fig. 1(a). There are 5 zones and a core peer exists in each zones which including member peers.

3.2 Constructing Overlay Network

When joining the P2P content sharing overlay network, a joining peer contacts a RP (Rendezvous Point) peer for bootstrapping. After obtaining list of connected peers, the joining peer sends a *connect* request with its NodeID to a peer which is closest to its NodeID among the list. When receives a *connect* request message, a peer acts according to its role (a core or a member in a zone). In case of member peer, it just forwards the message to the core peer within the same zone because only the core peer manages a zone. In case of a core peer, it compares NodeID in *connect* request message with range of its zone. If the NodeID does not belongs to the zone, the message is forwarded to a neighboring zone until reaching to the destination node according to message routing in Section 3.4. When the message reaches a core peer with a zone covering its NodeID, the core peer splits his zone while each zone owns their own NodeIDs. Zone splitting will be presented in Section 3.3. Every peers in P2P content sharing overlay network must hold a routing table: a list of neighboring zones. This routing table is used to route every messages until reaching the destination peer and maintain the overlay network. we define neighbors as zones which overlap each other's x or y coordinates. Therefore, after splitting a zone into two, change of the zone is known to their neighbors and to the members if there is at least one member peer within the zone. As the result of splitting the zone, new core takes a part of zone and hash table from old zone.

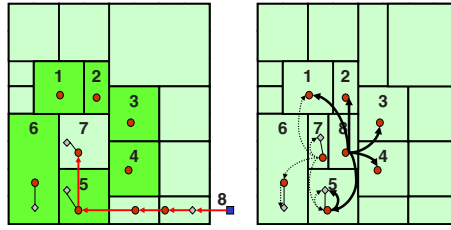


Fig. 2. (a) Connect request routing when peer 8 joins to zone 7 as zone 8. (b) After splitting zone 7 into zone 7 and 8.

3.3 Zone Management

Now, we describe the zone management which is used when new peers join in order to preserve the overlay network. When a joining peer's NodeID belongs to a zone, zone split is performed by the core peer in the zone. The zone is split in a direction of x or y axis according to the following.

To select the direction of split, a core peer calculates the absolute differences of x and y coordinate between their NodeIDs and then compares two absolute difference values. The zone is split along the axis of larger absolute difference. Border between two zones is chosen by mean value between two values of x or y coordinate according to direction of split zone. It is required that range of each zone should cover network identifiers of a core peer and member peers. After splitting a zone, neighbors of two zones are rearranged in order to maintain the P2P content sharing overlay network. For an example, neighbors of zone 7 in Fig. 2(a) are zone $\{1,2,3,4,5,6\}$. After splitting the zone 7 in Fig. 2(b), neighbors of zone 7 are zone $\{1,5,6,7,8\}$ and neighbors of zone 8 are zone $\{1,2,3,4,5,7\}$. The change from zone splitting should be informed to each zone's neighbors and then makes neighbors modify their own neighboring zones. Also, any change of a zone should be notified to member peers within the zone. Fig. 2(b) illustrates the notification of changes to its own member peers and neighboring zones. A joining peer's NodeID is the same with core peers's NodeID and the joining peer is added to as a member peer in the zone. A shared binary tree is constructed and maintained among member peers within a zone. This tree is used to deliver messages between member peers within a zone without any message duplication. Fig. 3 shows the shared binary tree between member peers and result after peer 6 joining and routing table of member peers within the same zone. This tree is constructed by taking advantage of order of member peers' IP addresses and stored in each peers. Update messages between members in a zone are passed along thick arrow lines in Fig. 3 when a peer joins and leaves.

3.4 Message Routing

Message routing is used to locate destination when inserting contents into and retrieving contents from P2P content sharing overlay network based on the

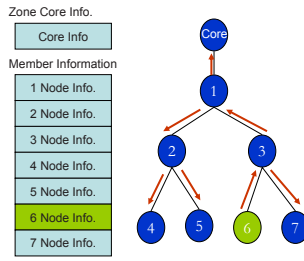


Fig. 3. Shared binary tree between member peers within a zone.

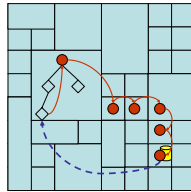


Fig. 4. Message routing in the proposed overlay network.

DHT. Our Approach defines two functions: $insert(key, value)$ and $value = retrieve(key)$ where $keys$ can be name of content and $value$ is an address of content owner. $Insert$ function provides means of distributing information of contents and $retrieve$ function means of getting information of actual content owner. Both $insert$ and $retrieve$ function use hash function with a key to get the (x, y) pair coordinate on the space. With this (x, y) pair coordinate, a message (i.e, insert and retrieve) is routed to the destination. Message routing in our approach is similar to that of CAN. Fig. 4 illustrates message routing when retrieving contents. To get a certain content, a node uses the $retrieve$ function with key . In $retrieve$ function, hash function is performed to get x and y coordinate. Message with this (x, y) coordinate is sent to a core peer directly. If the core peer does not manage a zone containing (x, y) , it just forwards to a neighbor zone which is closer to x . After reaching x , core peers select a neighbor zone whose range contains x and closer to y as next routing hop. When reaching the destination of $retrieve$ message, a core peer send $value$ to the requesting peer.

4 Experiment and Results

We use GT-ITM to create network topology with 10,000 nodes and experiment with randomly selected nodes among them. Simulator which is based on myns [7] is implemented to measure performance of message routing and verify our approach. To simplify our experiment, we use 2 bytes IP addresses so that x and y coordinate is ranged from 0 to $2^8 - 1$. Average number of neighboring zones in each zone are shown in Fig. 5. Each node maintains a routing table for their

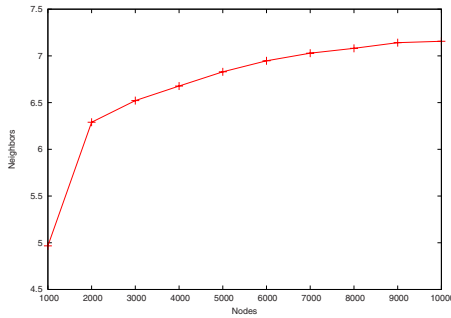


Fig. 5. Average number of neighbors.

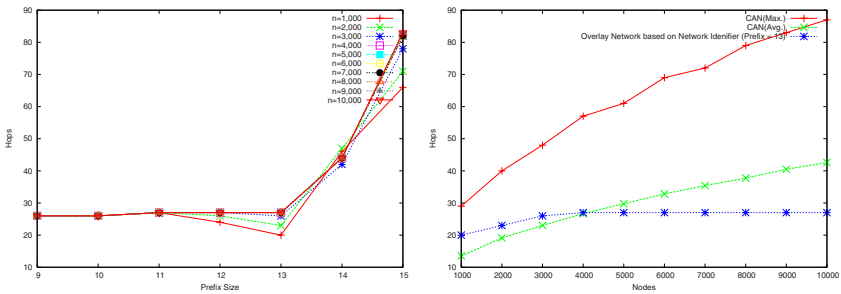


Fig. 6. (a)The number of message routing hops when changing length of network identifiers. (b) Comparison of the network identifier based overlay network with CAN.

neighbor zones’ information such as ranges of neighbor zones, IP addresses and port numbers of neighbor core nodes. The result shows average neighboring zones are about 7 even though the number of nodes increases. Therefore, it can be realized that our proposal P2P network is scalable.

Fig. 6(a) shows the effect of grouping as decreasing the length of network identifiers. The number of prefix means IP masks in Classless Inter-Domain Routing. Therefore, decreasing the number of prefix means that the number of nodes having the same identifier increase and more nodes are grouped into a zone. When the number of prefix is 15 in Fig. 6(a), the number of message routing hops are more than 67 regardless of the number of participating nodes. This means that most of nodes occupy their own zone and there are many zones involving in message routing. The result shows that the number of message routing hops are declining to 28 hops as the number of prefix decreases. When the number of prefix is less than 13, there is no gain. It shows the number of prefix, 13, is enough for grouping nodes. Therefore, we prove that grouping nodes with network identifiers help improving the performance of P2P content sharing overlay network. Fig. 6(b) shows the performance of network identifier based overlay network with 13 prefix is better than CAN as the number of nodes increases.

5 Conclusion

In this paper, we showed that grouping by network identifiers for P2P content sharing overlay network based on DHT improves the performance of message routing. Also, our P2P content sharing overlay network is scalable even though the number of nodes increases by average number of neighbors. Nodes on the same local network become a group so that physical network topology was reflected on our P2P network. In the future, we have to solve limitation of grouping by network identifiers because reducing the length of network identifier is not easy.

Acknowledgement. This work was supported in part by MIC through RBRC at GIST and in part by MOE through BK21.

References

1. S. Androutsellis-Theotokis and D. Spinellis, "A survey of peer-to-peer file sharing technologies," *Athens Univ. of Economics and Business*, 2002.
2. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content-addressable network," in *Proc. ACM SIGCOMM*, 2001.
3. V. N. Padmanabhan, H. J. Wang, and P. A. Chou, "Resilient peer-to-peer streaming," in *Proc. IEEE ICNP 2003*, Nov. 2003.
4. M. Hefeeda, A. Habib, B. Botev, D. Xu, and B. Bhargava, "Promise: peer-to-peer media streaming using Collectcast," in *Proc. ACM Multimedia 2003*, Nov. 2003.
5. Z. Xu and Z. Zhang, "Building low-maintenance expressways for P2P systems", *Hewlett-Packard Labs: Palo Alto. HPL-2002-41*, 2002.
6. Z. Xu, C. Tang, and Z. Zhang, "Building topology-aware overlays using global soft-state", in *Proc. The 23rd International Conference on Distributed Computing Systems*, May 2003.
7. S. Banerjee, myns, <http://www.cs.umd.edu/~suman>.

A Forward-Backward Voice Packet Loss Concealment Algorithm for Multimedia over IP Network Services

Mi Suk Lee¹, Hong Kook Kim², Seung Ho Choi³, Eung Don Lee¹, and Do Young Kim¹

¹ Electronics and Telecommunications Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Korea

{lms, edlee, dyk}@etri.re.kr

² Gwangju Institute of Science and Technology, 1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea

hongkook@gist.ac.kr

³ Seoul National University of Technology, 172 Gongreung 2-dong, Nowon-gu, Seoul 139-743, Korea

shchoi@snut.ac.kr

Abstract. In this paper, we propose a voice packet loss concealment algorithm in order to improve voice quality for both multimedia over IP and voice over IP services. The proposed algorithm estimates the coding parameters of lost frames by combining forward and backward prediction from the good frames before and after the lost frames. The performance of the proposed algorithm is evaluated on the ITU-T G.729 coder, and it is compared with the performance of the conventional algorithms in terms of objective and subjective quality measures. From the PESQ score comparison and the listening test, it is found that the proposed algorithm provides better voice quality than the conventional ones.

1 Introduction

Multimedia over Internet protocol (MoIP) applications, such as IP telephony and video streaming, continue to gain popularity. In MoIP systems, one or several encoded video or audio data are grouped into a packet for the transmission through packet networks. Fig. 1 shows the multimedia data, typically video and/or audio (voice) data, transmission in a packet network. The packet network for most MoIP systems operate based on RTP/UDP/IP, but they do not have any quality of service (QoS) control mechanism [1]. Thus, packet losses could occur due to network congestion. A packet loss is also declared when the packet has not been arrived yet within the delay time of a playout buffer on the receiver side. When a packet loss rate exceeds a given threshold, received video or audio (voice) becomes unintelligible. To reduce the quality degradation caused by packet losses, there have several approaches been developed that are categorized by adaptive multimedia [2], QoS control in the Internet [3], forward

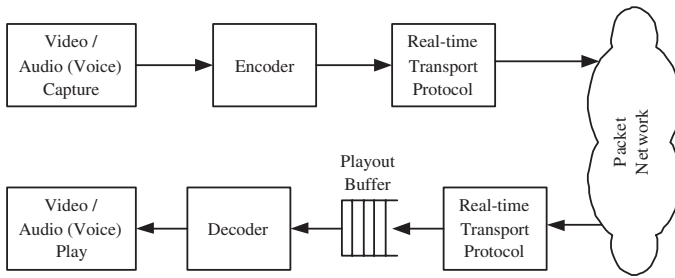


Fig. 1. Multimedia data transmission over packet network

error correction and packet loss concealment [4][5], and partial packet discard in asynchronous transfer mode (ATM) networks [6].

In this paper, we propose a voice (speech) packet loss concealment (PLC) algorithm for improving voice quality in both MoIP and VoIP systems. As shown in Fig. 1, the playout buffer for reducing the effects caused by delay jitter is an essential component of all MoIP and VoIP receivers, and it plays a main role in proposing a PLC algorithm in this paper. By assuming that the size of the playout buffer is enough to store at least one future good frame, we can utilize this good frame to improve the performance under a packet (frame) loss¹ condition without any extra delay. In fact, this assumption can be accepted at most of time [7].

In order to show the performance improvement of the proposed PLC algorithm over the conventional PLCs, we choose ITU-T G.729 [8] because it is a standard speech coder for H.261 and H.323 VoIP protocols. Moreover, G.729 is based on the code-excited linear prediction (CELP) analysis-by-synthesis algorithm, and thus we can apply the proposed technique to other VoIP standard speech coders including ITU-T G.723.1 [9] and ITU-T G.728 [10].

The PLC algorithms can be classified into the sender-based algorithm and the receiver-based algorithm with regard to the place where the concealment algorithm works. The sender-based algorithms, e.g., forward error correction (FEC), are more effective than receiver-based algorithms but require additional bits used for being processed in the decoder when frame losses occur [11]. On the other hand, the receiver-based algorithms including the repetition based forward PLC [8] and the interpolative PLC [12] have advantages over the sender-based algorithms since they do not need any additional bits, and thus we can use the already existing standard speech encoders without any modification.

In this paper we propose a receiver-based forward-backward PLC algorithm for MoIP and VoIP, which utilizes a future good frame as well as a previous good

¹ A packet loss results in more than one frame losses depending on the packet size, but a packet loss concealment algorithm is realized by repeating a frame loss concealment algorithm as many times as the packet size. Therefore, the terminologies of packet and frame appear with the same meaning in a view of a concealment algorithm.

frame in order to reconstruct the lost frames. The performance of the proposed algorithm is evaluated on the G.729 coder under different packet size and frame erasure rate conditions, and it is compared with those of the conventional ones.

Following this introduction, we briefly review two conventional receiver-based PLC algorithms in Section 2. And then, we propose a forward-backward prediction based PLC algorithm in Section 3. In Section 4, the performance of the proposed algorithm is demonstrated by using the objective and subjective speech quality measures. Also, we compare the performance of the proposed algorithm to that of the conventional ones. Finally, we summarize our results in Section 5.

2 Receiver-Based PLC Algorithms

In this section, we review two types of conventional receiver-based PLC algorithms. One is a forward concealment algorithm adopted in most standard speech coders. The other is based on an interpolative scheme to estimate the parameters of the lost frames by using a future good frame.

2.1 Forward Concealment Algorithm

In the forward PLC algorithms, the parameters of the lost current frame are estimated by extrapolating those of the previous good frame. That is, the parameters of the lost frame estimated by repeating the down-scaled version of the previous ones [8]. The PLC algorithm used in G.729 that is widely used for VoIP belongs to this category. The specific steps taken for reconstructing a lost frame in G.729 are 1) repeating the synthesis filter parameters, 2) attenuating adaptive and fixed codebook gains followed by attenuating the memory values of the gain predictor, and 3) randomly generating the excitation. This approach works well under wireless communication environments where the delay is an essential issue so there is no time to wait for the future good frames in the receiver.

2.2 Interpolative Concealment Algorithm

An interpolative PLC algorithm was proposed by assuming that in a VoIP system a future good frame is available in the playout buffer just after a series of lost frames [12]. Thus, this interpolative PLC algorithm could reconstruct a lost frame by interpolating the parameters of the previous and future good frames. However, it has been applied only to estimate the adaptive codebook parameters of G.729, while the other parameters including line spectral frequencies (LSF), fixed codebook gains and indices were obtained by using the forward PLC algorithm described in the previous subsection. Nevertheless, the interpolative PLC algorithm gave better speech quality than the forward PLC algorithm [12].

3 Proposed Forward-Backward PLC Algorithm

The CELP-based encoders represent speech signal as spectral envelope and excitation, and then quantize them for the transmission. The low bit-rate speech

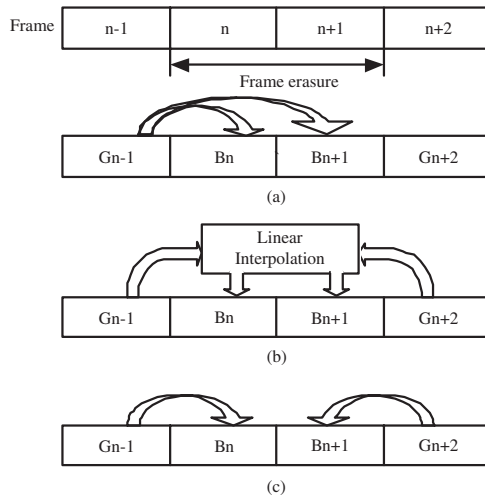


Fig. 2. Procedure of each PLC algorithm; (a) forward PLC, (b) interpolative PLC, and (c) forward-backward PLC algorithms, where B and G mean a bad (lost) frame and a good frame, respectively

coders are able to achieve toll-quality performance by exploiting the correlation between adjacent analysis frames when quantizing the coding parameters. By incorporating this property into PLC processing, we propose a forward-backward PLC (FB-PLC) algorithm. Fig. 2 demonstrates the basic idea of three PLC algorithms such as forward PLC (F-PLC), interpolative PLC (I-PLC), and FB-PLC. In the figure, the n -th and $(n+1)$ -th frames are lost. Therefore, a PLC algorithm is needed to reconstruct the n -th and $(n+1)$ -th frames. As shown in Fig. 2(a), the F-PLC algorithm approximates the parameters of the n -th and $(n+1)$ -th frames by repeating the parameters of the $(n-1)$ -th frame with down-scaling. The performance of F-PLC depends on the degree of correlation between the previous good frame and the lost frame. Thus, if the frames are consecutively erased, the performance degradation of F-PLC is related to the duration of frame loss.

On the other hand, I-PLC utilizes the information of the future good frame stored in playout buffer as shown in Fig. 2(b), where the parameters of the n -th and $(n+1)$ -th frames are estimated by linearly interpolating those of the $(n-1)$ -th and $(n+2)$ -th frames. I-PLC provides better performance than F-PLC, but it has difficulties in estimating the parameters of the lost frames, especially when the lost frames are from the transient regions of speech.

To overcome the problem of I-PLC, the proposed FB-PLC algorithm estimates the parameters of the lost frames from the most adjacent frame among the previous and the future good frames because adjacent frames have higher correlation than the two frames apart. Fig. 2(c) shows an example of the FB-PLC procedure. The n -th frame is reconstructed from the $(n-1)$ -th frame (forward

concealment) because the $(n-1)$ -th frame is closer to the n -th frame than the $(n+2)$ -th frame. Similarly, the $(n+1)$ -th frame is reconstructed from the $(n+2)$ -th frame (backward concealment). The procedure of estimating the parameters of the n -th frame by using the forward prediction in FB-PLC is different from that in F-PLC since the estimates of parameters in FB-PLC are bounded by the result of the backward prediction. Similarly, the estimates of the $(n+1)$ -th frame parameters are obtained by the backward prediction and bounded by the result of the forward prediction.

In G.729, the proposed FB-PLC algorithm is implemented in a frame basis for LSF, but in a subframe basis for pitch and codebook parameters. The parameters of a lost frame are estimated by forward and backward prediction based on the parameters of the good frame closest to the lost frame. When the number of lost frames is even, each half of the frames are estimated by using the previous or the future good frames close to each lost frame. On the other hand, when it is odd, the LSF parameters for the center frame in the lost frames are approximated by averaging the parameters estimated by the forward and backward prediction from the previous and future good frames, respectively. However, we do not have such a problem in estimating pitch and codebook parameters because the number of subframes in the lost frames is equally divided by half even if the number of the lost frames is odd.

4 Performance Evaluation

We applied FB-PLC, F-PLC, and a modified version of I-PLC into G.729, and compared the performance of the three concealment algorithms in terms of the objective and subjective quality. Even though the I-PLC algorithm was implemented only for the pitch parameter as in [12], we extended the interpolation procedure to the estimation process of LSF and codebook gain. It turned out to be that this approach gave the better performance than the original form of the I-PLC algorithm. From here on, this approach is referred to as modified I-PLC (MI-PLC).

To evaluate the objective and subjective quality of the PLC algorithms, we prepared 8 Korean sentences spoken by 8 speakers (4 males and 4 females) from the NTT-AT database [13]. Each sentence was 8 second long and was sampled at 16 kHz, followed by down-sampling to 8 kHz using the ITU-T G.191 software tool [14].

First, we compared the performance of each PLC algorithm in a waveform point of view. It was assumed that the three consecutive frames of the original speech has been lost during the transmission as shown in Fig. 3(a). Figs. 3(b), (c), and (d) show the reconstructed speech signals by G.729 employing F-PLC, MI-PLC, and FB-PLC, respectively. It is shown from the figure that F-PLC and MI-PLC fail to reconstruct the speech segments corresponding to the lost frames, but FB-PLC could at least generate a periodic waveform, which is much better than the waveforms processed by F-PLC and MI-PLC. In addition, MI-PLC gave better performance in a view of reconstructed waveform than F-PLC

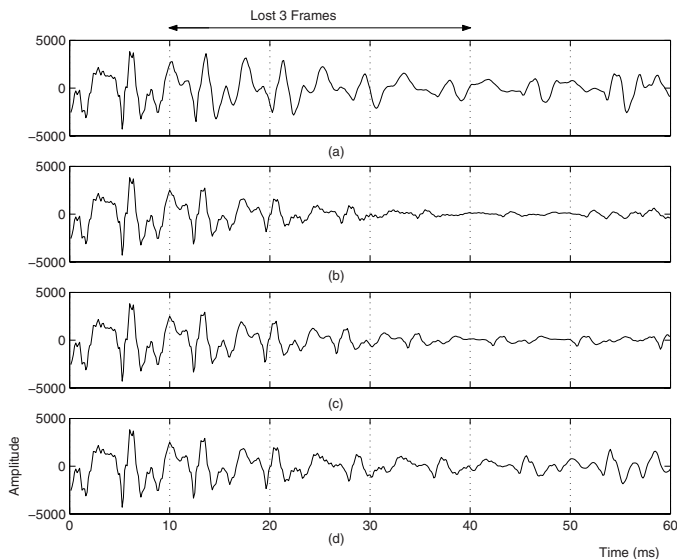


Fig. 3. Performance comparison of each PLC algorithm in a waveform point of view; (a) original speech signal and speech signals processed by (b) F-PLC, (c) MI-PLC, (d) FB-PLC, respectively

but worse than FB-PLC. It is generally expected that MI-PLC works well for stationary regions but does not for transition regions because MI-PLC is based on the linear interpolation of the previous and future good frames. Therefore, MI-PLC could not effectively estimate the parameters of the lost frames if the parameters of the good frames before the lost frames are significantly different from those after the lost frames as shown in Fig. 3.

Next, ITU-T P.862 (PESQ) [15] was used as an objective quality measure for each PLC algorithm and the results were summarized in Table 1. The PESQ score for a CELP-based coder was known to be correlated with the perceptual quality measured by mean opinion score (MOS) [15]. In this experiment, we modeled the VoIP channel as a randomly erased channel with a different frame erasure rate (FER) from 1% to 15%. In order to evaluate the robustness of PLC algorithms to the burst frame erasure, the packet size was set to 10, 20, and 30 ms. Table 1 shows the PESQ scores of the conventional and proposed PLC algorithms. As discussed in Section 2, MI-PLC achieved better PESQ scores than F-PLC. However, the FB-PLC algorithm gave the best performance under all FER conditions. What is interesting in the table is that the PESQ score is not always degraded as the packet size increases [16], especially when the FER is low. In addition, it is important to give an emphasis on the point that the proposed FB-PLC achieved significantly better performance for larger packet size and higher FER than F-PLC and MI-PLC.

Table 1. PESQ scores for each PLC algorithm

Packet Size (ms)	PLC Type	Frame Erasure Rate (%)					
		1	3	5	7	10	15
10	F-PLC	3.67	3.38	3.24	3.09	2.91	2.72
	MI-PLC	3.69	3.46	3.32	3.20	3.04	2.84
	FB-PLC	3.71	3.50	3.39	3.28	3.13	2.92
20	F-PLC	3.62	3.27	3.09	2.93	2.70	2.52
	MI-PLC	3.65	3.35	3.17	3.03	2.81	2.63
	FB-PLC	3.68	3.42	3.30	3.16	3.02	2.78
30	F-PLC	3.62	3.30	2.99	2.81	2.60	2.32
	MI-PLC	3.65	3.42	3.13	2.97	2.80	2.52
	FB-PLC	3.70	3.47	3.21	3.09	2.93	2.74

Table 2. Preference test results between the MI-PLC algorithm and the proposed FB-PLC algorithm at 5% FER

Speaker	Preference Score (%)	
	MI-PLC	FB-PLC
Female	33.33	66.67
Male	30.56	69.44
Average	31.95	68.05

Finally, we performed AB-preference tests between MI-PLC and FB-PLC at a frame erasure of 5%. To process the speech sentences with a more general frame erasure pattern, the error insertion device in G.191 was used. In the experiments, the burstiness of the channel was set to 0.7, and also the maximum number of consecutive frame loss was restricted to three. These processed four female and four male sentence pairs were presented to 9 listeners at a randomized order. Table 2 shows the relative preference of FB-PLC to MI-PLC. The proposed FB-PLC algorithm was significantly preferred over the MI-PLC algorithm.

5 Conclusion

In MoIP systems, if a packet loss rate exceeds a given threshold, received video and audio (voice) become unintelligible. Thus, several algorithms have been studied to improve the quality of received video and audio (voice). In this paper, we propose an improved voice packet loss concealment (PLC) algorithm for CELP-based speech coders that are widely used for voice compression in MoIP or VoIP services. The proposed algorithm estimates the coding parameters of the lost frames by combining forward and backward prediction from the good frames before and after the erased frames. We evaluate the performance of the proposed algorithm on the ITU-T G.729 under the various simulated channel environments, and compare it with those of the forward PLC and interpolative PLC algorithms. From the PESQ measure and the listening test, it was shown that

the proposed PLC algorithm gave better PESQ scores than others and more than two-thirds of the participants to the listening test preferred the proposed PLC algorithm.

Acknowledgement. This work was sponsored in part by the Ministry of Information and Communication (MIC) of Korea through the R&D Project for Advanced IT Core Technology and also in part by MIC through the University IT Research Center Project.

References

1. Hersent, O., Gurle, D., Petit, J.-P.: IP telephony: Packet-based multimedia communications systems. Addison Wesley (2000)
2. Atiquzzaman, M., Hassan, M. (eds.): Adaptive real-time multimedia transmission over packet switching networks. Real-Time Imaging, (2001) 219-220
3. Wroclawski, J.: The use of RSVP with IETF integrated services. RFC2210, Sept. (1997)
4. Bolot, J.-C., Fosse-Parisis, S., Towsley, D.: Adaptive FEC-based error control for Internet Telephony. Proc. of IEEE INFOCOM, vol. 3, Mar. (1999) 1453-1460
5. Perkins, C., Hodson, O., Hardman, V.: A Survey of packet loss recovery techniques for streaming audio. IEEE Network, vol. 12, no. 5, Sept.-Oct. (1998) 40-48
6. Armitage, G.J., Adams, K.M.: Packet reassembly during cell loss. IEEE Network, vol. 7, no. 5, Sept. (1993) 26-34
7. Kim, M.J., Lee, K.H., Kwon, C.H.: A dynamic redundant audio transmission in VoIP systems. IEICE Trans. Commun., vol. E86-B, no. 6, June (2003) 2056-2059
8. ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). ITU, Mar. (1996)
9. ITU-T Recommendation G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. ITU, Mar. (1996)
10. ITU-T Recommendation G.728: Coding of speech at 16 kbit/s using low-delay code excited linear prediction, ITU, Oct. (1992)
11. Wah, B.W., Su, X., Lin, D.: A survey of error-concealment schemes for real-time audio and video transmissions over the Internet. Proc. IEEE Int'l Symposium on Multimedia Software Engineering, Taipei, Taiwan, Dec. (2000) 17-24
12. de Martin, J.C., Unno, T., Viswanathan, V.: Improved frame erasure concealment for CELP-based coders. Proc. ICASSP, Istanbul, Turkey, June (2000) 1483-1486
13. NTT-AT: Multi-lingual speech database for telephony. (1994)
14. ITU-T Recommendation G.191: Software tools for speech and audio coding standardization. ITU, Nov. (2000)
15. ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU, Feb. (2001)
16. Kim, H.K., Kang, H.-G.: A frame erasure concealment algorithm based on gain parameter re-estimation for CELP coders. IEEE Signal Processing Letters, vol.8, no. 9, Sept. (2001) 252-256

A Delay-Based End-to-End Congestion Avoidance Scheme for Multimedia Networks

Li Yan, Bin Qiu, and Lichang Che

School of Computer Science and Software Engineering,
Monash University, Vic 3800, Australia.
{lyan,bq,carski}@csse.monash.edu.au

Abstract. Multimedia applications are emerging on the current Internet. They have stringent requirements for quality of service (QoS) in terms of loss ratio, delay, delay jitter and throughput. Loss-based congestion control algorithms in the current Internet, however, are not effective for delivering multimedia applications. As a solution, a delay-based end-to-end congestion avoidance (DECA) algorithm is proposed. In this scheme, a proportional controller is designed to control the sending rate. The design principle of the controller is presented in detail in this paper. Simulation results provide evidence that the proposed scheme is capable of improving network utilization and reducing loss ratios, mean delay and delay jitter. Its superiority over other end-to-end congestion control algorithms is also demonstrated.

Keywords: Delay-based congestion control, Congestion control, DECA.

1 Introduction

In the Internet, real-time applications are gradually growing and occupying a greater proportion of traffic. Transmission of real-time applications makes demands on bandwidth, delay, and loss requirements. However, current transmission technology only provides reliable service for non-real-time applications and is less effective in providing required service for real-time applications. For example, the heuristically-based TCP transport protocol is not suitable for delivering real-time applications, since its self-caused drop mechanism causes a higher loss ratio and larger delay and decreased throughput, and the congestion window is frequently cut in half leading to instability of video quality. Therefore, emerging real-time applications are mainly delivered via UDP. UDP's lack of congestion control, however, can cause network instability and a fairness issue when UDP traffic competes with TCP traffic. Thus, the evolution of the end-to-end congestion control should, ideally, be able to improve overall network performance including all QoS parameters for all types of traffic.

The purpose for TCP-friendly congestion control mechanism is to smoothly adjust the sending rate, while making traffics compatible with TCP traffics. A well-known example is TFRC[1], which deploys a TCP throughput equation to calculate its control point. It is noted that all TCP-friendly congestion control

mechanisms are heuristically-based, but the use of a loss event as a congestion signal does not reduce packet loss. Therefore, more proactive congestion detection and avoidance methods are highly desirable.

A lot of research efforts have been spent on delay-based congestion avoidance mechanisms, e.g. [2][3][4]. The best known method among these mechanisms is TCP/Vegas [4], in which a proactive congestion avoidance method was used to compare the sending rate with the expected rate. This expected rate is calculated based on total packet size sent during a measured period and minimum round-trip time. The congestion window is updated according to the difference between the actual sending rate and the expected rate. This scheme achieves a lower packet loss ratio and increased throughput in some cases but it can not reduce packet loss when there are a large number of flows competing in a limited buffer capacity [4][5].

Other delay-based TCP congestion control schemes include Novel TCP [6] and FAST TCP [7]. They are designed to improve utilization in high-speed networks. It should be noticed that the above delay-based congestion control mechanisms are also window-based, and therefore may experience burst flow and causes queue build-up. A rate-based and delay-based end-to-end congestion control mechanism was designed by Morita [8] to improve overall network performance. However, this scheme only works well in rather simple networks where the bottleneck link is shared by a very small number of users. Morita warns that if control parameters are set inappropriately, then the control system can not work correctly even in simple networks [8].

All the preceding schemes improve one or two QoS parameters, but not the overall network performance. In this paper, the proposed scheme (DECA) is presented which focuses on improving overall network performance in order to provide better QoS in delivering all types of applications over multimedia networks.

The DECA involves a delay-based end-to-end congestion avoidance method, which uses a controller [9] to adjust the sending rate for maintaining network node buffer size within a suitable range. The proposed scheme also deploys rate-based method rather than window-based method for avoiding burst flow. An important aspect of this scheme is that it is completely end-to-end because it does not rely on QoS functionality in routers, explicitly congestion notification (ECN), and still uses first-in first-out buffer management.

The rest of the paper is organized as follows: Section 2 presents the proposed congestion avoidance mechanism; Section 3 presents the performance evaluation; and the paper is concluded in Section 4.

2 Proposed Scheme

The presented scheme draws particularly on end-to-end delay-based congestion control schemes described in Section 1. It solves congestion problems from a control theoretical perspective. Therefore, four key aspects of this scheme are

discussed: network model, selection of control point, control frequency, and the proposed rate adjustment algorithm.

2.1 Network Model

A network model is built to define the relationship between buffer size and delay. This modeling method has been thoroughly described in previous work [3]. A network consists of a group of hosts, routers/switches and links. The round-trip time (rtt) mainly contains queuing time, processing time and propagating time. The total processing time and propagation time is the minimum time indicated as rtt_{min} for a packet transmitting along a connection path. The queuing time is reflected by the occupied proportion of buffer. To simplify this analysis, it is supposed that the queuing time is only spent on the bottleneck router along the forward path and all packets have the same size. If the bottleneck router has the speed C and its maximum buffer size is B , then the maximum delay time is $(\frac{B}{C} + rtt_{min})$. From the above description, the occupied buffer size is estimated according to Formula(1).

$$Buff = (rtt - rtt_{min}) * C. \quad (1)$$

2.2 Choice of Set-Point and Control Frequency

The objective of the DECA control is to maintain buffer size of the bottleneck node within a suitable range to improve utilization, reduce loss ratio and mean delay. A congestion control scheme[10] for ATM networks suggests that half the buffer size as a control point balances the loss ratio and the degree of utilization. In the scheme to the DECA, as well, half the buffer size was chosen as a control point in order to improve network utilization and reduce loss ratio.

The use of individual measurement of delay to determine if the sending rate should be increased or decreased is not suitable for the Internet because of highly variable delay [11]. Short control intervals cause unnecessary oscillations. In contrast, long control intervals can not adapt to the dynamic network state. According to system control theory, optimal control frequency depends on feedback delay, which is the time between applying the change and getting feedback from the network corresponding to the change [2]. Therefore, the average round-trip is selected as the control interval. The average control interval is shown as follows:

$$t_{arg} = (1 - \beta)t_{arg}(n - 1) + \beta t(n). \quad (2)$$

Where β is the filter gain $0 < \beta < 1$, which is suggested as 0.2. $t(n)$ is the average round-trip time during the n th control interval.

2.3 Rate Adjustment Algorithm

The rate adjustment process in this proposed scheme is divided into two parts: slow start and congestion avoidance. Once the slow start process commences,

the source sends packets at a predefined minimum rate and then that rate is increased by one packet size per control interval as shown in Formula(3). If a packet loss event has not happened, then the source can keep increasing the rate until maximum rate is reached. When a packet loss is detected, the sending rate is decreased by one packet size during the next control interval as indicated in Formula(4), and the slow start is terminated and the congestion avoidance process commences.

$$Rate_{n+1} = Rate_n + \frac{packetSize}{t_{n+1}}. \tag{3}$$

$$Rate_{n+1} = Rate_n - \frac{packetSize}{t_{n+1}}. \tag{4}$$

The purpose of the slow start process is that a new connection can quickly probe the maximum delay to seek a fair control point. A lost packet implies a maximum delay met by either its own connection or other connections. If all connections have opportunities to experience the same maximum queuing time, then control points for all connections are the same, and fairness can be guaranteed among all connections.

During the congestion avoidance process, the controller was used to update the sending rate, as follows:

$$Rate_{n+1} = Rate_n + \frac{k_p Err_n}{t_{n+1}} \tag{5}$$

where the $Rate_n$ is the current sending rate, the k_p is the proportional gain, and the control error Err_n indicates the distance between the measured buffer size and reference value. t_{n+1} indicates the next control point of time.

The reference value is half maximum buffer size, whereas the measured buffer size is the maximum buffer size during the current interval. This choice of using the maximum delay is more conservative than using average delay in order to prevent packets from loss. The Err_n is indicated as follows:

$$Err_n = Buf_{fref} - Buf_n = (\lambda(rtt_{max} - rtt_{min}) - (rtt_n - rtt_{min})) * C, \tag{6}$$

where λ is the set-point described in Section 2.2 as 0.5; rtt_{max} is the maximum delay; rtt_{min} is the minimum delay; rtt_n is the maximum delay during current control interval; and C is the bottleneck capacity. If the maximum queuing delay is larger than the control point, then the source reduces the sending rate, otherwise the source increases the sending rate.

To define the k_p , we assumed a scenario where there were n flows in total. If all flows experience the same buffer size, then the k_p should be $\frac{1}{n}$. Given the lack of information on the number of competing flows, k_p will be decided according to the current flow's contribution to bottleneck capacity as follows:

$$Rate_{n+1} = Rate_n + \frac{Rate_n}{C} * \frac{Err_n}{t_{n+1}} \tag{7}$$

$$= Rate_n + \frac{Rate_n(\lambda(rtt_{max} - rtt_{min}) - (rtt_n - rtt_{min}))}{t_{n+1}}. \tag{8}$$

The Formula(8) above represents the rate adjustment algorithm where the sending rate is adjusted according to the current sending rate and the difference between measured buffer size and the target buffer size. If there is no acknowledgment received during a control interval, the sending rate is decreased by one packet size per control interval as in Formula(4).

3 Simulation Results

The DECA algorithm was implemented on the Linux platform and evaluations performed with the *ns2* network simulator[12]. The simulation results provided us with evidence that the DECA mechanism is capable of supporting an effective congestion avoidance mechanism, which significantly improves overall network performance. In this section, the simulation results are discussed.

In the simulation a dumbbell topology consists of a bottleneck link with *n* sources and destinations connected. The bottleneck link is 10Mb/s with a propagation delay of 10ms. The links at source end and destination end have the capacity of 100Mb/s. It is assumed that each source can send at a rate exceeding bottleneck speed. Routers in the bottleneck use the FIFO drop mechanism. The buffer size is set at 10 packets with a mean size of 1000 bytes per packet. The start time for each connection is set for a little later than the previous connection by 0.1s, and the initial transmission rate of each connection is 0.1Mb/s. The simulation runs 150 s in total. The DECA is then compared with existing congestion control mechanisms including loss-based window-based, loss-based rate-based and delay-based window-based mechanisms represented, respectively, by TCP/Sack, TFRC, and TCP/Vegas.

3.1 Static Scenario

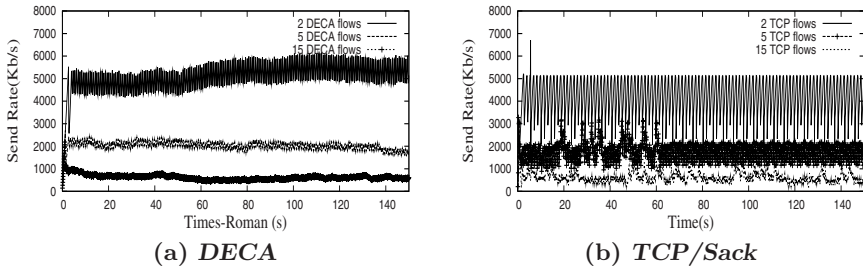
Tabs.1 and 2 demonstrate the performance comparison between different mechanisms. There were 20 simulation results for each mechanism. Each simulation ran a different number of connections.

Table 1. Utilization and loss ratio comparison with different mechanisms

Number of flows	TCP/Sack		TFRC		TCP/Vegas		DECA	
	utilization%	loss ratio%	utilization%	loss ratio%	utilization%	loss ratio%	utilization%	loss ratio%
1	67	0.43	99	19.19	98	0	98	1.81
4	85	0.47	100	0.44	98	0	98	0.84
8	87	1.44	100	1.85	100	12.05	98	0.34
10	88	1.96	100	2.97	100	24.46	98	0.26
14	89	3.13	100	6.04	100	14.68	98	0.2
18	89	4.25	100	8.64	100	18.87	98	0.2
20	89	4.8	100	9.38	100	19.24	98	0.21

Table 2. Queue length comparison with different mechanisms

Number of flows	TCP/Sack		TFRC		TCP/Vegas		DECA	
	length	std	length	std	length	std	length	std
1	2.69	2.54	6.72	2.68	2.00	0.1	6.21	2.47
4	3.69	2.74	6.06	1.65	6.98	0.36	4.51	2.65
8	4.88	3.13	6.49	1.22	6.85	0.88	3.52	2.24
10	4.93	2.84	6.62	1.13	6.77	1.43	3.31	2.28
14	5.53	2.75	6.74	1.25	7.79	0.75	3.06	1.79
18	6.02	2.62	6.83	1.33	7.42	0.83	3.08	1.75
20	6.13	2.59	6.87	1.34	7.84	0.74	3.12	1.71

**Fig. 1.** Sending rate

In the Tab.1, it can be seen that the DECA, the TCP/Vegas and the TFRC can achieve a higher link utilization (above 98%) regardless the number of connections. In contrast, link utilization by deploying the TCP/Sack depends on the number of connections. The more the TCP flows, the higher the utilization. Obviously, TCP/Sack results in lower utilization than other mechanisms under the same condition.

Meanwhile, Tab.1 shows the comparison of loss ratio. It can be seen that the TCP/Vegas mechanism can completely avoid packet loss when the number of flows is less than 4, but it results in much higher loss ratio than the other mechanisms when the number of flows is greater than 4. The TFRC also produces higher loss ratios than the TCP/Sack. In contrast, the DECA loss ratio remains low regardless the number of flows. It is noticed that when the number of the DECA flow is 1, the loss ratio for the DECA is higher than TCP/Sack loss ratio. However, it appears better to obtain higher utilization by suffering a slightly higher loss ratio than lower utilization as experienced by TCP/Sack traffic.

Tab.2 presents the mean queue length with its standard deviation using different mechanisms. The queue length with its standard deviation reflects the queuing delay with its variation. The unit of the queue length is the average number of packets. It can be seen that the mean queue length using the DECA is convergent to a stable point around 3 along with an increased number of flows, while the standard deviation gradually decreases from 2.5 to around 1.7.

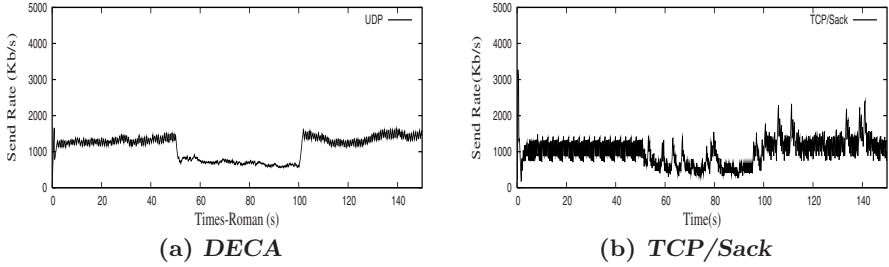


Fig. 2. Sending rate

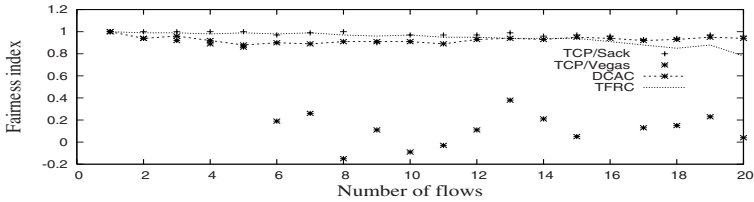


Fig. 3. Fairness behavior

In contrast, the mean queue length for TCP/Sack gradually increases when the number of flows is increased. As a comparison, the DECA mechanism performs better in terms of mean queue length, whereas the TFRC and the TCP/Vegas behave better in reducing queue length variation.

Fig.1 displays the variation of sending rate when using the DECA and the TCP/Sack, with 2, 5, and 15 flows respectively competing in the bottleneck link. The TCP/Vegas and the TFRC had a stable sending rate in simulation results. Because of page limits, the results are not presented here. In Fig.1(a), oscillations by the DECA tend to be reduced when the number of flows is increased. Compared with the TCP/Sack flow (Fig.1(b)), the DECA flow shows obviously lower oscillation than the TCP/Sack flow.

3.2 Dynamic Scenario

Fig.2 displays how one flow responds to a change of available bandwidth by deploying different mechanisms. From 0s to 50s, 8 flows were running, at 50s another 7 flows started to run, and at 100s, 7 flows exited from the network at the same time. In Fig.2(b), the TCP/Sack flow reaches their new state in a slow and unstable way, whereas the DECA flow naturally and quickly scales with the change of available bandwidth as indicated in Fig.2(a). In the simulation results, both of the TFRC flow and the TCP/Vegas flow slowly respond to the change of bandwidth than the DECA flow. As a result, the slower speed responding to the reduced available bandwidth would lead to an increase in loss ratio, whereas slower speed responding to increased available bandwidth would cause a decrease in utilization.

3.3 Fairness Analysis

Fig.3 displays the fairness behavior with the four mechanisms. The fairness index is calculated by using average sending rate with the standard deviation between connections. The fairness index for the TCP/Sack, the DECA, and the TFRC demonstrate the similar behaviors. The TCP/Vegas mechanism has poor fairness behavior when there are more connections. However, there also exists a fairness issue for the DECA, when connections have larger difference in propagation time.

4 Conclusion

The proposed DECA scheme is capable of enhancing overall network performance including effectively avoiding packet loss, improving network utilization and reducing mean delay, and delay jitter. Future enhancement of the scheme such as the use of a fuzzy controller [13] may improve the stability of sending rate and further reduce delay jitter.

References

1. S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation-based congestion control for unicast applications," in Proc. ACM Sigcomm, pp. 43-56, 2000.
2. R. Jain, "A delay-based approach for congestion avoidance in interconnected heterogeneous computer networks," *Comput.Commun.Rev.*, vol. 19,no 5, pp. 56-71, 1989.
3. Z. Wang and J. Crowcroft, "Eliminating periodic packet losses in the 4.3-Taboe BSD TCP congestion control algorithms," *Comput.Commun.Rev.*, vol. 22,no.2, pp. 9-16, 1992.
4. L. Brakmo and L. Peterson, "TCP Vegas: end to end congestion avoidance on a global Internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1465-1480, 1995.
5. J. Martin, A. Nisson, and I. Rhee, "Delay-based congestion avoidance for TCP," *IEEE/ACM Transactions on networking*, vol. 11,no.3, 2003.
6. W. Xu, A. G. Quresh, and K. W. Sarkies, "Novel TCP congestion control scheme and its performance evaluation," *IEE Proc.Commun.*, vol. 149,No.4, 2002.
7. C. Jin, D. Wei, and S. H. Low, "FAST TCP: from theory to experiment," <http://netlab.caltech.edu/pub/papers/fast-030401.pdf>, 2003.
8. M. Morita, H. Ohsaki, and M. Murata, "Designing a delay-based adaptive congestion control mechanism using control theory and system identification for TCP/IP network," in Proc. ITCOM, 2002.
9. T. K. Kiong, W. Q. Guo, H. C. Chieh, and T. J. Hagglund, *Advances in PID Control*: Springer, 1999.
10. S. Keshav, "A control-theoretic approach to flow control," In Proc. Communications architecture and protocols, pp. 3-15, 1993.
11. V. Paxson, "Measurements and analysis of end-to-end Internet Dynamics," in *Information and computer science*. Berkeley: California, 1997.
12. "The network simulator -ns2." <http://www.isi.edu/nsnam/ns/>.
13. B. Qiu, "A predictive fuzzy logic congestion avoidance scheme," in Proc. IEEE Globecom, pp. 967-971, 1997.

Dynamic Bandwidth Allocation for Internet Telephony*

Yiu-Wing Leung

Department of Computer Science, Hong Kong Baptist University, Kowloon Tong,
Hong Kong. ywleung@comp.hkbu.edu.hk.

Abstract. Internet telephony is promising for long-distance calls because of its low service charge and value-added functions. To provide Internet telephony to the general public, a service provider can operate a telephone gateway in each servicing city to bridge the local telephone network and the Internet, so that users can use telephones or fax machines to access this gateway for services. In this paper, we propose a dynamic bandwidth allocation scheme for two purposes: (1) each telephone gateway can fully utilize the available bandwidth to serve more telephone and fax sessions and (2) it can respond to the changing environments. We exploit three properties for dynamic bandwidth allocation. First, in a telephone session, each user usually alternates between speaking and listening. When a user is not speaking, she does not send any voice stream and hence the bandwidth can be dynamically released from this session for the other sessions. Second, voice traffic is elastic because it can be further compressed at the cost of a lower quality. Third, fax traffic is flexible because it can be temporarily delayed. We exploit these three properties to allocate bandwidth to telephone and fax sessions dynamically. When a telephone gateway adopts dynamic bandwidth allocation, it can serve more telephone and fax sessions while providing acceptably good quality-of-service (QoS), and it can give more stable QoS when the available bandwidth varies.

Keywords: Internet telephony, dynamic bandwidth allocation.

1 Introduction

Internet telephony is a telephony service through the Internet [1]. It is promising for long-distance calls because of its low service charge and value-added functions. Some market analyses also reveal its enormous potential (e.g., see [2]).

To use Internet telephony, we can use a computer connected to the Internet. However, many people of the general public are not Internet users (e.g., only 6.6% of population in China are Internet users [3]), and some Internet users may not be able to access computers or the Internet at certain time. To provide Internet telephony to all users, a service provider can operate a *telephone gateway* in each servicing city to bridge the local telephone network and the Internet [4], so that users can use telephones or fax machines to access this gateway for long-distance telephone services (see Fig. 1).

* This project is supported by the RGC Grant HKBU 2092/01E.

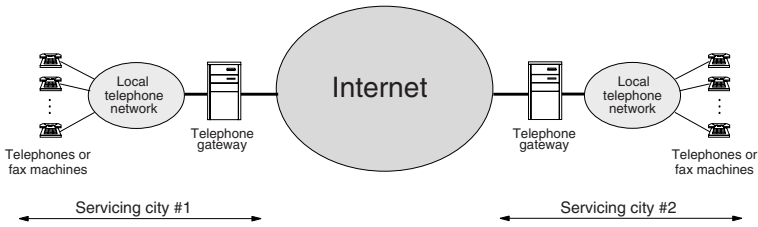


Fig. 1. In each servicing city, a telephone gateway is used to bridge the local telephone network and the Internet.

Each gateway collects telephone and fax traffic from its servicing city, and sends it to the destination gateway through the Internet. Each gateway should properly allocate bandwidth to the telephone and fax sessions for two reasons:

1. Each gateway should fully utilize the available bandwidth to serve more sessions while providing acceptably good quality-of-service (QoS). Then the service provider gets more revenue or the service charge is lower.
2. Each gateway has to handle several types of changes. First, the available bandwidth between two gateways may change with time because the other traffic in the Internet may compete for this bandwidth. Second, the bandwidth requirement of each telephone session is time-varying because each user usually alternates between speaking and listening. Third, the number of telephone and fax sessions may change with time because some ongoing sessions may be completed and some new sessions may be admitted.

In this paper, we propose a dynamic bandwidth allocation scheme for Internet telephony. This scheme can significantly increase the number of telephone and fax sessions that can be served, and it can give more stable QoS when the available bandwidth varies.

2 Principle of Dynamic Bandwidth Allocation

Each telephone gateway sends traffic to and receives traffic from the remote telephone gateway through the Internet. Without loss of generality, we consider the outgoing bandwidth of each gateway. The available bandwidth may change with time, so the gateway measures this quantity regularly (say, once per five minutes). This measurement can be done by many effective measurement techniques, such as passive [5] or active measurement techniques [6]. In particular, passive measurement techniques are effective for this application. It is because the gateways transmit telephone and fax traffic to each other continuously, so they can utilize the information in these transmissions for measurement.

We use admission control to control the number of ongoing telephone and fax sessions. After admitting a session, we must serve it until it is completed. To allocate the available bandwidth to the admitted sessions, we exploit the following three properties:

1. In a telephone session, a user alternates between speaking and listening. When she is speaking, she sends a voice stream and the session is *active*; otherwise, she does not send any voice stream and the session is *inactive*. The gateway can use a voice activity detector [7] to detect whether a session is active, and it does not allocate any bandwidth to the inactive session. In this manner, the available bandwidth can be better utilized.
2. Voice traffic is elastic because it can be further compressed at the cost of a lower quality (e.g., by discarding the less important bits). Let a voice stream require: (i) a *full bandwidth* of b before further compression and (ii) a *partial bandwidth* of b' after further compression. When the available bandwidth is not enough (e.g., it becomes smaller or many telephone sessions become active), we allocate partial bandwidth to some active telephone sessions so that these sessions can still send voice streams.
3. Fax traffic is flexible because it can be temporarily delayed. When there is spare bandwidth, we allocate more bandwidth to it for faster transmission; when there is temporary congestion, we allocate less bandwidth to it.

When there is congestion (e.g., the available bandwidth becomes smaller or many telephone sessions become active), a session may be temporarily allocated a small bandwidth. To ensure acceptably good QoS for telephone and fax sessions, we adopt the following QoS measures and requirements:

1. When we cannot allocate any bandwidth to an active telephone session, this session is interrupted until we can allocate bandwidth to it. The probability that an active telephone session is interrupted is called **probability of session interruption** T_{SI} . We ensure that T_{SI} is not larger than an acceptable value T_{SI}^* (e.g., $T_{SI} \leq 0.001$).
2. It is desirable to allocate full bandwidth to every active telephone session. The probability that an active telephone session is allocated full bandwidth is called **probability of full bandwidth allocation** T_{FBA} . We ensure that T_{FBA} is at least equal to an acceptable value T_{FBA}^* (e.g., $T_{FBA} \geq 0.99$).
3. It is desirable to allocate more bandwidth to every fax session for faster transmission. The average bandwidth allocated to a fax session is called **average allocated bandwidth** F_{AAB} . We ensure that F_{AAB} is at least equal to an acceptable value F_{AAB}^* (e.g., $F_{AAB} \geq 10$ kbps).

Based on the above discussion, we adopt the following principle for dynamic bandwidth allocation:

1. We measure the available bandwidth regularly (e.g., once per five minutes).
2. We execute admission control to control the number of ongoing telephone and fax sessions while fulfilling the QoS requirements (i.e., $T_{SI} \leq T_{SI}^*$, $T_{FBA} \geq T_{FBA}^*$ and $F_{AAB} \geq F_{AAB}^*$).
3. For telephone sessions, we try to allocate full bandwidth to all the active sessions, and do not allocate any bandwidth to the inactive sessions. If the available bandwidth is not enough, we try to allocate full bandwidth to the largest number of active sessions and partial bandwidth to the remaining active sessions. If the available bandwidth is still not enough, we allocate partial bandwidth to the largest number of active sessions.

4. For fax sessions, we allocate the remaining bandwidth to them. When there is temporary congestion, the fax sessions may be allocated a small bandwidth. Nevertheless, when some telephone sessions are completed or become inactive, we can allocate more bandwidth to the fax sessions again.

3 Analysis

In this section, we analyze the bandwidth allocation and the QoS measures. With these analytical results, we can perform dynamic bandwidth allocation in section 4. We consider the outgoing bandwidth of any telephone gateway. Let the available bandwidth be B , there be n_t and n_f telephone and fax sessions respectively, and α be the probability that a telephone session is active.

3.1 Bandwidth Allocation

Based on the principle described in section 2, we analyze the amount of bandwidth allocated to telephone and fax sessions in the following. Among the n_t telephone sessions, let j of them be active. We distinguish four cases:

- *Case 1.1:* $j \leq \lfloor \frac{B}{b} \rfloor$. We allocate full bandwidth to all the j active telephone sessions, and allocate the remaining bandwidth $B - jb$ to the fax sessions.
- *Case 1.2:* $\lfloor \frac{B}{b} \rfloor < j \leq 1 + \lfloor \frac{B-b}{b'} \rfloor$. We allocate full bandwidth and partial bandwidth to m_1 and m_2 active telephone sessions respectively, where m_1 and m_2 are determined as follows. The total bandwidth allocated to the active telephone sessions cannot be larger than the available bandwidth, so $m_1 b + m_2 b' \leq B$. Moreover, we allocate either full or partial bandwidth to each active session, so $m_1 + m_2 = j$. Furthermore, we allocate full bandwidth to the largest number of active telephone sessions, so m_1 is as large as possible. Based on these properties, we get $m_1 = \lfloor \frac{B - j b'}{b - b'} \rfloor$ and $m_2 = j - \lfloor \frac{B - j b'}{b - b'} \rfloor$. After allocating bandwidth to the active telephone sessions, we allocate the remaining bandwidth to the fax sessions.
- *Case 1.3:* $1 + \lfloor \frac{B-b}{b'} \rfloor < j \leq \lfloor \frac{B}{b'} \rfloor$. We allocate partial bandwidth to all the j active telephone sessions because $B \geq j b'$, and allocate the remaining bandwidth $B - j b'$ to the fax sessions.
- *Case 1.4:* $j > \lfloor \frac{B}{b'} \rfloor$. We allocate partial bandwidth to $\lfloor \frac{B}{b'} \rfloor$ active telephone sessions and the remaining bandwidth $B - \lfloor \frac{B}{b'} \rfloor b'$ to the fax sessions.

3.2 Probability of Session Interruption T_{SI}

To derive T_{SI} , we distinguish two cases:

- *Case 2.1:* $n_t \leq \lfloor \frac{B}{b'} \rfloor$. Even when all the n_t telephone sessions are active, we can still allocate either full or partial bandwidth to every active session (see Cases 1.1, 1.2 and 1.3). Therefore, $T_{SI} = 0$.

- *Case 2.2:* $n_t \geq \lfloor \frac{B}{b'} \rfloor + 1$. Let j of the n_t telephone sessions be active, where j follows the binomial distribution:

$$\Gamma(j) \equiv \binom{n_t}{j} \alpha^j (1 - \alpha)^{n_t - j} \quad (1)$$

If $j \leq \lfloor \frac{B}{b'} \rfloor$, every active session is allocated either full or partial bandwidth (see Cases 1.1, 1.2 and 1.3), so no active session is interrupted. If $j > \lfloor \frac{B}{b'} \rfloor$, only $\lfloor \frac{B}{b'} \rfloor$ active telephone sessions are allocated partial bandwidth and the other $j - \lfloor \frac{B}{b'} \rfloor$ active sessions are temporarily interrupted (see Case 1.4), so the conditional probability that an active telephone session is interrupted is equal to $(j - \lfloor \frac{B}{b'} \rfloor) / j$. Therefore, $T_{SI} = \sum_{j=\lfloor \frac{B}{b'} \rfloor + 1}^{n_t} \Gamma(j) \cdot (j - \lfloor \frac{B}{b'} \rfloor) / j$.

3.3 Probability of Full Bandwidth Allocation T_{FBA}

To derive T_{FBA} , we distinguish two cases:

- *Case 3.1:* $n_t \leq \lfloor \frac{B}{b'} \rfloor$. Even when all the telephone sessions are active, they are allocated full bandwidth (see Case 1.1) and hence $T_{FBA} = 1$.
- *Case 3.2:* $n_t > \lfloor \frac{B}{b'} \rfloor$. Among the n_t telephone sessions, let j of them be active. When $j \geq 1$, the probability distribution of j is given by: $\frac{1}{1 - (1 - \alpha)^{n_t}} \cdot \Gamma(j)$. Given any value of j , the conditional probability that an active session is allocated full bandwidth can be determined as follows:
 - If $1 \leq j \leq \lfloor \frac{B}{b'} \rfloor$, every active telephone session is allocated full bandwidth (see Case 1.1), so the conditional probability is equal to one.
 - If $\lfloor \frac{B}{b'} \rfloor < j \leq 1 + \lfloor \frac{B-b}{b'} \rfloor$, only $\lfloor \frac{B-jb'}{b-b'} \rfloor$ active telephone sessions are allocated full bandwidth (see Case 1.2), so the conditional probability is equal to $\lfloor \frac{B-jb'}{b-b'} \rfloor / j$.
 - If $j > 1 + \lfloor \frac{B-b}{b'} \rfloor$, no active telephone session is allocated full bandwidth (see Cases 1.3 and 1.4), so the conditional probability is equal to zero.

We remove the condition on j to obtain T_{FBA} as follows:

$$T_{FBA} = \frac{1}{1 - (1 - \alpha)^{n_t}} \left\{ \sum_{j=1}^{\lfloor \frac{B}{b'} \rfloor} \Gamma(j) + \sum_{j=\lfloor \frac{B}{b'} \rfloor + 1}^{\min(n_t, 1 + \lfloor \frac{B-b}{b'} \rfloor)} \Gamma(j) \cdot \left(\frac{\lfloor \frac{B-jb'}{b-b'} \rfloor}{j} \right) \right\} \quad (2)$$

3.4 Average Allocated Bandwidth F_{AAB}

We let j of the n_t telephone sessions be active and distinguish four cases:

- *Case 4.1:* $n_t \leq \lfloor \frac{B}{b'} \rfloor$. Based on Case 1.1, F_{AAB} is:

$$F_{AAB} = \frac{1}{n_f} \sum_{j=0}^{n_t} \Gamma(j) \times (B - jb) \quad \text{for } n_t \leq \lfloor \frac{B}{b'} \rfloor \quad (3)$$

– *Case 4.2:* $\lfloor \frac{B}{b} \rfloor < n_t \leq 1 + \lfloor \frac{B-b}{b'} \rfloor$. Based on Cases 1.1 - 1.2, F_{AAB} is:

$$F_{AAB} = \frac{1}{n_f} \left\{ \sum_{j=0}^{\lfloor \frac{B}{b} \rfloor} \Gamma(j) \cdot (B - jb) + \sum_{j=\lfloor \frac{B}{b} \rfloor + 1}^{n_t} \Gamma(j) \cdot \left(B - \lfloor \frac{B-jb'}{b-b'} \rfloor b - \left(j - \lfloor \frac{B-jb'}{b-b'} \rfloor \right) b' \right) \right\} \quad (4)$$

– *Case 4.3:* $1 + \lfloor \frac{B-b}{b'} \rfloor < n_t \leq \lfloor \frac{B}{b} \rfloor$. Based on Cases 1.1 - 1.3, F_{AAB} is:

$$F_{AAB} = \frac{1}{n_f} \left\{ \sum_{j=0}^{\lfloor \frac{B}{b} \rfloor} \Gamma(j) \cdot (B - jb) + \sum_{j=\lfloor \frac{B}{b} \rfloor + 1}^{1 + \lfloor \frac{B-b}{b'} \rfloor} \Gamma(j) \cdot \left(B - \lfloor \frac{B-jb'}{b-b'} \rfloor b - \left(j - \lfloor \frac{B-jb'}{b-b'} \rfloor \right) b' \right) \right. \\ \left. + \sum_{j=2 + \lfloor \frac{B-b}{b'} \rfloor}^{n_t} \Gamma(j) \cdot (B - jb') \right\} \quad (5)$$

– *Case 4.4:* $n_t > \lfloor \frac{B}{b'} \rfloor$. Based on Cases 1.1 - 1.4, F_{AAB} is:

$$F_{AAB} = \frac{1}{n_f} \left\{ \sum_{j=0}^{\lfloor \frac{B}{b} \rfloor} \Gamma(j) \cdot (B - jb) + \sum_{j=\lfloor \frac{B}{b} \rfloor + 1}^{1 + \lfloor \frac{B-b}{b'} \rfloor} \Gamma(j) \cdot \left(B - \lfloor \frac{B-jb'}{b-b'} \rfloor b - \left(j - \lfloor \frac{B-jb'}{b-b'} \rfloor \right) b' \right) \right. \\ \left. + \sum_{j=2 + \lfloor \frac{B-b}{b'} \rfloor}^{\lfloor \frac{B}{b'} \rfloor} \Gamma(j) \cdot (B - jb') + \sum_{j=\lfloor \frac{B}{b'} \rfloor + 1}^{n_t} \Gamma(j) \cdot (B - \lfloor \frac{B}{b'} \rfloor b') \right\} \quad (6)$$

4 Dynamic Bandwidth Allocation

Each gateway measures the available bandwidth B from itself to the remote gateway regularly (e.g., once per five minutes). When B or the number of active telephone sessions is changed, the gateway allocates bandwidth to the telephone sessions based on the results derived in Cases 1.1 - 1.4, and then allocates the remaining bandwidth to the fax sessions.

When a user requests to set up a new telephone session, the gateway decides whether it can admit this session. A new telephone session would affect the bandwidth allocation to the existing telephone and fax sessions, so it would affect all the three QoS measures. If the gateway admits this new telephone session and all the three QoS requirements can still be fulfilled (where the three QoS measures can be computed based on the results derived in Cases 2.1 - 2.2, 3.1 - 3.2 and 4.1 - 4.4), the gateway admits this session and allocates bandwidth accordingly; otherwise, it rejects the request.

When a user requests to set up a new fax session, the gateway decides whether it can admit this session. A new fax session would affect the bandwidth allocation to the existing fax sessions, but it would not affect the bandwidth allocation to the existing telephone sessions. Therefore, it would only affect the QoS measure for fax session (i.e., F_{AAB}). If the gateway admits this new fax session and the QoS requirement can still be fulfilled (where F_{AAB} can be computed based on the results derived in Cases 4.1 - 4.4), the gateway admits this session and allocates bandwidth accordingly; otherwise, it rejects the request.

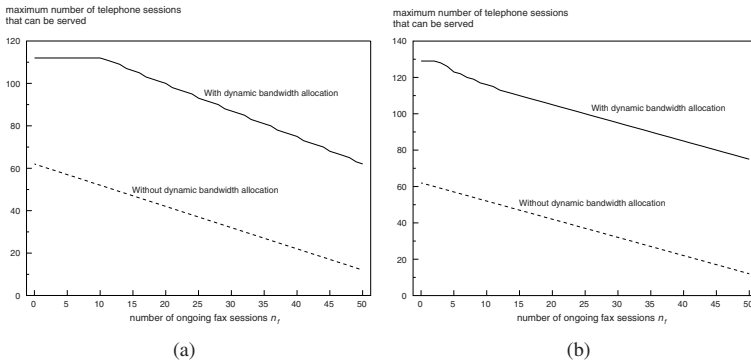


Fig. 2. Advantage of dynamic bandwidth allocation: (a) tighter QoS requirements, (b) looser QoS requirements.

Table 1. The resulting QoS when the available bandwidth is reduced from 1 Mbps to 0.75 Mbps, 0.5 Mbps, 0.25 Mbps at time instances 1, 2, 3 respectively.

Time instances	Available bandwidth	Probability of session interruption T_{SI}		Probability of full bandwidth allocation T_{FBA}		Average allocated bandwidth per fax session F_{AAB}	
		With DBA	Without DBA	With DBA	Without DBA	With DBA	Without DBA
0	1.00 Mbps	0.000	0.000	1.000	1.000	34.0kbps	16.0kbps
1	0.75 Mbps	0.000	0.000	1.000	1.000	21.5kbps	0.7kbps
2	0.50 Mbps	0.000	0.225	1.000	0.775	9.0kbps	0.20kbps
3	0.25 Mbps	3.603×10^{-6}	0.625	0.575	0.375	0.19kbps	0.50kbps

5 Numerical Results

In this section, we present two numerical examples to illustrate the advantages of dynamic bandwidth allocation (DBA). We adopt the following parameter values: $B = 1$ Mbps, $b = 16$ kbps, $b' = 8$ kbps, and $\alpha = 0.5$.

Example 1: In this example, we demonstrate that DBA can significantly increase the number of sessions that can be served. We let the QoS requirements be $T_{SI} \leq 0.001$, $T_{FBA} \geq 0.99$ and $F_{AAB} \geq 10$ kbps. Fig. 2(a) shows the maximum number of telephone sessions that can be served as a function of n_f . When $n_f = 10$, the gateway can serve only 52 telephone sessions without DBA, but it can serve 112 telephone sessions with DBA (i.e., the percentage increase is 115%). When $n_f = 50$, the gateway can serve only 12 telephone sessions without DBA, but it can serve 62 telephone sessions with DBA (i.e., the percentage increase is 416%). These results demonstrate that DBA can better utilize the available bandwidth to serve more sessions while providing acceptably good QoS. When the QoS requirements are looser and become $T_{SI} \leq 0.001$, $T_{FBA} \geq 0.90$ and $F_{AAB} \geq 8$ kbps, Fig. 2(b) shows that DBA can further increase the maximum number of sessions that can be served. For example, when $n_f = 50$, the

gateway can serve only 12 telephone sessions without DBA but it can serve 75 telephone sessions with DBA. With looser QoS requirements, more sessions can share the available bandwidth to a larger extent.

Example 2: In this example, we demonstrate that DBA can result in more stable QoS when the available bandwidth varies. Suppose the available bandwidth is 1 Mbps, and it is reduced to 0.75 Mbps, 0.5 Mbps, 0.25 Mbps at time instances 1, 2, 3 respectively. Table 1 shows the resulting QoS values at these time instances. In general, when DBA is adopted, the three QoS measures are relatively less affected and hence the resulting QoS is more stable. For example, consider the case in which the available bandwidth is reduced from 1 Mbps to 0.5 Mbps. When DBA is adopted, the probability of session interruption T_{SI} remains to be 0, the probability of full bandwidth allocation T_{FBA} remains to be 1, and the average allocated bandwidth for each fax session F_{AAB} becomes 9 kbps. When DBA is not adopted, however, T_{SI} is increased from 0 to 0.225, T_{FBA} is decreased from 1 to 0.775, and F_{AAB} becomes only 0.2 kbps.

6 Conclusions

We considered the Internet telephony systems which use telephone gateways to bridge the local telephone networks and the Internet, such that each user can use a telephone or a fax machine to access a telephone gateway for long-distance telephone services. We proposed a dynamic bandwidth allocation scheme for two purposes: (1) each telephone gateway can fully utilize the available bandwidth to serve more telephone and fax sessions and (2) it can respond to the changing environments. We demonstrated that when a telephone gateway adopts dynamic bandwidth allocation, it can serve significantly more telephone and fax sessions while providing acceptably good QoS, and it can give more stable QoS when the available bandwidth varies.

References

1. *IEEE Internet Computing*, Special Issue on Internet Telephony, vol. 6, no. 3, May/June 2002.
2. M. Winther, "Web talk 2000: market forecast and analysis," *International Data Corporation Report #W22019*, April 2000.
3. J. Lyman, "Internet users in China number nearly 80 million," *ECT News Network*, 15 Jan 2004.
(Available at <http://www.ecommercetimes.com/perl/story/32610.html>.)
4. M. Gaynor, "Linking market uncertainty to VoIP service architectures," *IEEE Internet Comput.*, vol. 7, no. 4, pp. 16-22, July-Aug. 2003.
5. M. Stemm, R. H. Katz, and S. Seshan, "A network measurement architecture for adaptive applications," *Proc. IEEE INFOCOM*, pp. 285-294, March 2000.
6. N. Hu and P. Steenkiste, "Evaluation and characterization of available bandwidth probing techniques," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 6, pp. 879-894, August 2003.
7. S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498-505, Sept. 2003.

A Broadcasting Technique for Holographic 3D Movie Using Network Streaming

Kunihiko Takano¹, Koki Sato², Ryoji Wakabayashi¹,
Kenji Muto¹, and Kazuo Shimada¹

¹ Tokyo Metropolitan College of Aeronautical Engineering, 8-52-1, Minami-senju,
Arakawa-ku, Tokyo, 116-8523, Japan

² Shonan Institute of Technology 1-1-25, Nishikaigan, Tsujido, Fujisawa-shi,
Kanagawa, 251-8511, Japan

Abstract. Lately, many kinds of transmitting techniques over the high speed communicating networks called broad-band internet have widespread among the general public. However, as for the transmission of 3D visual images of moving objects, it seems to be not so well-known. In this paper, we present a transmitting process of 3D visual moving objects adopted network streaming technique and the hologram in which various information concerning 3D objects are recorded as fringe patterns. Applying this method, an excellent transmission is shown to be achieved for 3D visual images of moving objects. Moreover, a good reconstruction of holographic images can be performed by the transmitted streaming data. From this result, it seems to be possible to develop a new transmitting process of 3D moving data utilizing well-known conventional techniques in this field.

1 Introduction

With the aid of recent progress in LAN and internet network systems, audio and visual datas have come to be transmitted with very high quality. In addition, a new broadcasting system[1] and a distant-learning system[2] have rapidly been put to practical use. When the datas are accumulated beforehand by the server (on demand), viewers and listeners are able to enjoy contents without any restriction of time and place. However, as for the transmission of 3D visual images of moving objects, there appears to be few reports asserting that the network streaming process can be extended to a transmitting technique and 3D visual images can be carried over the network using this method. In this paper, we study a transmitting technique for holographic 3D visual images of moving objects, and show that the network streaming technique can play an essential role in the transmission of holographic 3D visual images over the networks. The reasons why the holography is particularly studied here are the following: (1) In holography, the information describing 3D visual objects are processed into two dimensional fringe patterns, so that the modern transmitting system established for planar images can be effectively applied. (2) Hologram is transformed into a highly redundant signals of 3D visual images, and thus, even if a certain

deterioration occurs in quality of images in the transmitting system, the loss of information describing 3D visual images is considered to be small. In this experiment, a visual signal transmitting process is performed using the hologram obtained by computer aided composition by transforming it as a Real Media Form (based on MPEG-4), which is one of the techniques of Network streaming process. (on demand transmission) As this result, an excellent transmission of 3D visual images of moving objects is seen to be possible. Since broad-band internet system has now widespread among the general public, it seems to be possible for 3D visual images to be delivered to the general public over the network by applying this technique introduced above.

2 Transmission of 3D Visual Images of Moving Objects Using Network Streaming Technique

2.1 Outline of the System

In order to study the possibility of transmission of 3D visual images of moving objects, we prepared a transmitting system shown in Fig.1, and made an experiment of "on demand" transmission. The system is composed of RTSP server, client PC and 3D display system. 3D display system is constructed by Laser and single DMD panel. RTSP server is adopted because of the facts that the visual data of moving objects cannot be down loaded as a file of contents and so the server appears to have very excellent property on the point of security. In addition, as a software, "Windows Media 9", "Real Media" and "Quick Time Streaming Server" are well-known, however, in our study. "Real Media" processing is adopted. As a network is shown in Fig.1(a), we think, in the future, internet system may play its role, but here, judging from the security, we adopted LAN in our College as shown in Fig.1(b). Moreover, the signs (1)–(7) in the figure show the directions of data flow. In the network shown in Fig.1(b), packet data can be transmitted in the following way: (1)–(4): From the client PC, which receives the data, various services are required to RTSP server through proxy. (5), (6): Required data is transmitted through proxy from RTSP server. (7): Datas are received at the client PC, from which services have been required to serve. Here, since the connecting rate of RTSP server to network is 10Mbps LAN, our network system works with the rate of 10Mbps.

The transmission of 3D visual images of moving objects was preceded as follows: (1) Prepare the CGH data of visual images of moving objects as a Real Media form (320*240pixel). (2) Store the CGH data in the RTSP server(PC) installed "Herix Server Basic9.0" (3) Receive the CGH data at the client PC in which "Real One Player 9.0" is installed. (4) Present the received CGH data on the web-browser in SMIL script form without changing the data size. Every area except that of CGH data is performed as black area on the display. (5) Input the CGH data appeared on the web-browser as RGB signals to the display system which performs 3D images of moving objects. (6) Reconstruct 3D images by illuminating He-Ne Laser on the hologram plane (DMD panel). In this processing 3D images are observed as virtual images in the deep side of DMD panel [4].

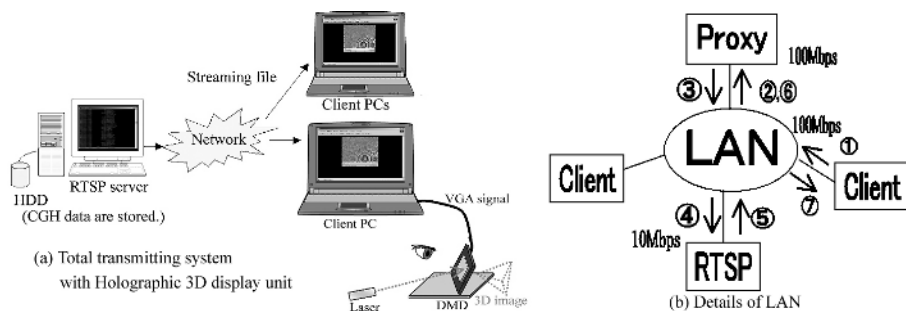


Fig. 1. Transmitting system of images

2.2 Specifications of PC Adopted in the Experiment

Specifications of RTSP server and client PC adopted in this experiment are shown in the table-1.

Table 1. Specifications of RTSP server

	RTSP Server	Client PC
CPU	Pentium II 450[MHz]	K6-III 400[MHz]
Memory	256[MB]	128[MB]
HDD	60[GB]	3.2[GB]
Graphic chip		MGA Millennium 8MB

2.3 Specifications of DMD Panel

In this experiment, a single DMD panel is used as a modulator which operates corresponding to the level of the signal to produce a hologram based on the CGH pattern. The specifications of DMD panel is given in table-2 made by Texas Instruments co.

Table 2. Specifications of RTSP server

Panel size	0.9[in](XGA)
Drive	Digital micro-mirror device
Refresh rate	360[Hz] (Max)
Contrast	420:1
Number of pixels	786,432(V1024×H768)
Pitch of pixels	17[μm]×17[μm]
Pixel size	16.2[μm]×16.2[μm]

3 Result

3.1 Estimation of Spatial Resolution Under CZP

Here, CGH is produced by adding CZP(=circular zone plate). Therefore, it may be very important to study suitable conditions of spatial resolution level required for the transmission of visual data of moving objects. We show some results obtained by applying CZP of real Media form, each of which is produced by changing the encoding rate. The setting of encoding rate is made by using target audience of Herix Producer9. In Fig.2, resulting several characteristics of spatial resolution are presented. Since the characteristic of spatial resolution was found to make notable changes at 20,50,100,350,750,1000kbps, we examined by choosing these values as encoding rates. As seen from Fig.2, high spatial resolutions are not attained at encoding rates 20,50 and 100kbps by the disturbance of aliasing, but encoding rates greater than 350kbps, comparatibly high spatial resolution is attained, and we can ignore the influence of aliasing. In fact, in such situations, highly faithfully reconstructed images have been obtained. On the other hand, encoding rate depends on the compressing ratio and PSNR(=Peak Signal to Noise Ratio). In Fig.3, the relations of compressing ratio vs. encoding rate and PSNR vs. encoding ratio are shown. Compressing ratio is given by the ratio between the amounts of data capacity of original CZP(non compressed AVI file) and that of the encoded file written in the Real Media form. In addition, PSNR[4] is defined by

$$PSNR = 20 \log_{10} \left(\frac{255}{\sqrt{MSE}} \right) \text{ dB} \quad (1)$$

were

$$MSE = \frac{\sum_{i,j} |F(i,j) - f(i,j)|^2}{x \cdot y} \quad (2)$$

where $F(i,j)$ and $f(i,j)$ are numbers of pixels at the point (i,j) in the reconstructed images from the output data and in the input images, respectively, and $x \cdot y$ denotes the total number of pixels. Fig.3 shows that compressing ratio changes linearly with respect to encoding rate, but PSNR saturates at nearly 700kbps of encoding rate. Further, noting Fig.3, we can see that highly faithful reconstruction is achieved when PSNR is set at nearly 31dB and encoding rate at 350kbps.

3.2 Characteristics of Reconstructed 3D Images by CGH

We shall consider some sufficient conditions to make it possible to perform a display of 3D visual images of moving objects recovered by CGH, which is transmitted over the network adopting network streaming technique. Let us take a triangular pyramid as an input image and produce an original CGH of this image. Examples of an input image ,CGH and a 3D image reconstructed with the help of this CGH are found in Fig.4. First, we produced a CGH of visual

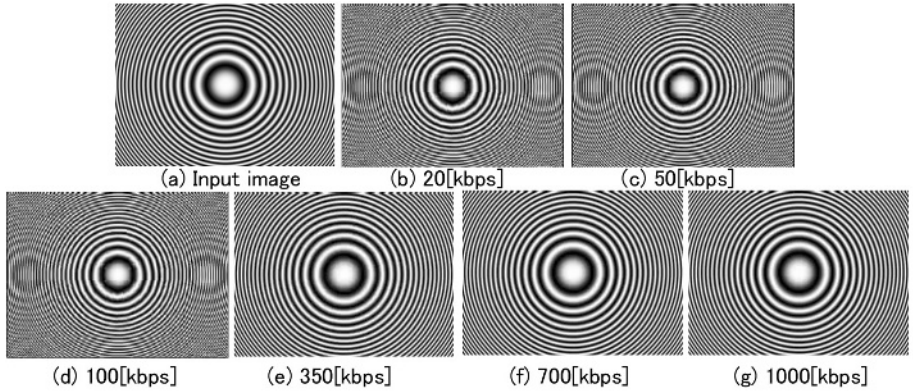


Fig. 2. Result of spatial frequency under C.Z.P

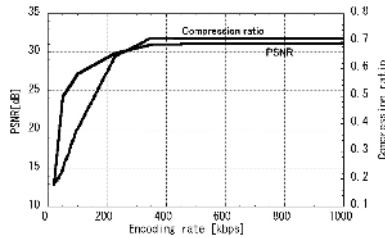


Fig. 3. Relations of compressing ratio and PSNR vs encoding rate

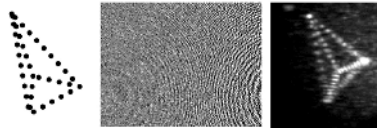


Fig. 4. CGH pattern and reconstructed image (Before encoding)

data of the image shown in Fig.4(b) without applying data compression, and transmitted it using the system in Fig.1. Then we performed a 3D visual image reconstruction and studied to find necessary conditions of the encoding rate to make a nice transmission of CGH over the network. Fig.5 gives relationships of compressing rate and PSNR vs. encoding rate.

Following results are obtained from Fig.5. 1) Both encoding rate and PSNR saturate at nearly 700kbps. 2) Even for 1500kbps, which is the maximum encoding rate, visual data is compressed with the ratio of 6Next, serving CGH data of Real Media form to RTSP, we tried to recover visual images of moving objects by client PC. Samples of the reconstructed images (camera:DCR-VX2000,effective

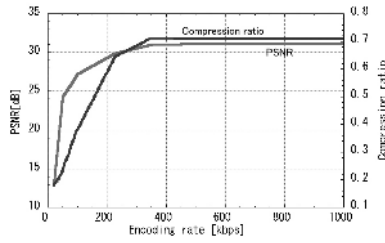


Fig. 5. Relationships of compressing ratio and PSNR vs. encoding rate for CGH

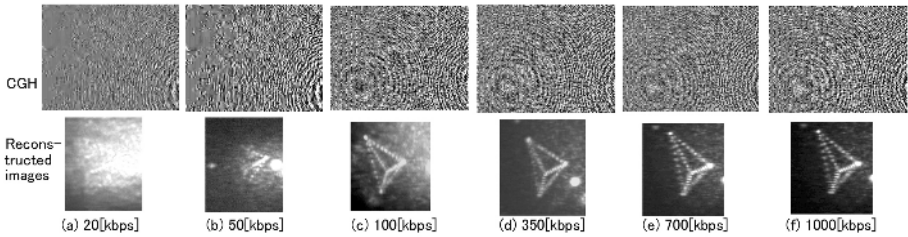


Fig. 6. CGH and reconstructed images for several encoding rates

pixel 380 thousand \times 3, Sony.co.Ltd.) are shown in Fig.6 together with the CGH under the encoding rate of 20,50,100,350,700 and 1000kbps. These encoding rates give typical characteristics of the reconstructed visual images. As this result, we can find following remarks: 1) At the encoding rate of 20kbps(nearly likely to 28.8k modem,PSNR=6.5dB) CGH pattern is disturbed by data compression, and 3D visual images are not obtained. 2) At the encoding rate of 50kbps(nearly likely to single ISDN,PSNR=7.5dB), background noise cannot be ignored, and 3D visual images are found with a little part of them missing. 3) For the encoding rate not smaller than 350kbps(PSNR=14dB) 3D visual images are reconstructed without any missed parts. It implies that in order to encode a CGH, its rate should be more than 350kbps. 4) At the encoding rate of more than 700kbps (PSNR=16dB) , background noise is suppressed to be small, and nice 3D visual images of moving objects are to be recovered. 5) Not so large differences cannot be found when the encoding rate is set at the value between 1000kbps and 700kbps. This shows that in order to make a nice transmission, it is desirable to choose the value of encoding rate around 700kbps(PSNR=16dB).

3.3 The Loss of Information Digits

Backgroundnoise is considered to result from the loss of information digits (lowering of PSNR), and diffracted angle of the wave seems to be smaller in proportion to the amount of lost information. Since, in our experiment, the image of a 3D moving picture is transmitted by decomposing it into 30 segments of

CGH patterns, each of which consists of 320×240 picture elements, the required bandwidth B for transmission of CGH is estimated as

$$B = 320 \times 240 \times 30 \times 24 = 55.3[\text{Mbps}]$$

on a general standard of uncompressed AVI file. It is based on the consideration which admits 24 bits color (3 Bytes) in the transmission of one picture element. In this sense, quite a high transmitting rate should be supported, and so, in order to construct a LAN transmitting system of the rate of 10 Mbps, data compression technique should be necessarily applied. Moreover, in holography, since a 3D image is reconstructed through reflection and diffraction of the wavefronts from DMD panel, the higher the spatial frequency of the wave becomes, the larger the diffracted angle of the wave tends to be, which enables us to reconstruct a 3D image in a larger size together with a wide visible region. Our system has adopted MPEG4 for compressing CGH data. It suggests that in lower rate data compression such as $100 \sim 350$ kbps, the loss of information in the wavefront from CGH results from a large lowering of spatial frequency, and in turn, it causes a deterioration in contrast and size of the images. On this view point, a relation between frequency and encoding rate is studied. In Fig.7, note that spatial resolving power W required for reconstruction of a 3D image is given by (3), and spatial frequency F , by (4) below.

$$W = \frac{\lambda}{\sin \theta} [m] \quad (3)$$

$$F = \frac{1}{W} [\ell p/m] \quad (4)$$

In the system, following settings are employed.

$$\lambda = 632.6 (nm); \quad \theta = 0.53 (deg); \quad W = 68.4 (\mu m); \quad F = 14.6 (\ell p/nm)$$

(W and F are determined by (3) and (4)).

Remak here that when CGH is employed, W corresponds to the value of $2 \times$ (pitch of picture element).

We investigated a 3D image reconstruction by changing the pitch of picture element of CGH from $19(\mu m)$ to $72(\mu m)$. Experimental samples are obtained on several typical points belonging to the interval $[19, 72](nm)$. They are shown in Fig.8.

From Fig.8, we see that partial deficiency is observed in reconstructed images when the pitch of a picture element is greater than $39(\mu m)$ and image reconstruction becomes impossible at $72(\mu m)$. Thus, when CGH's are encoded with 50(kbps) and 20(kbps) by MPEG4, they are regarded as the ones constructed under a larger pitches of picture element than $39(\mu m)$ and $72(\mu m)$, respectively. In addition, if CGH is encoded with a rate greater than 700(kbps), the spatial frequency needed for reconstruction is considered to be well-supplied and, as a result, the size of a 3D image and its visible region are properly obtained. These results are considered to be due to a highly redundant property of the hologram.

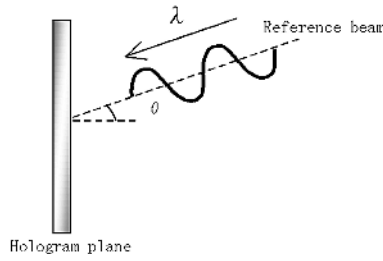


Fig. 7. Principle of spatial resolution

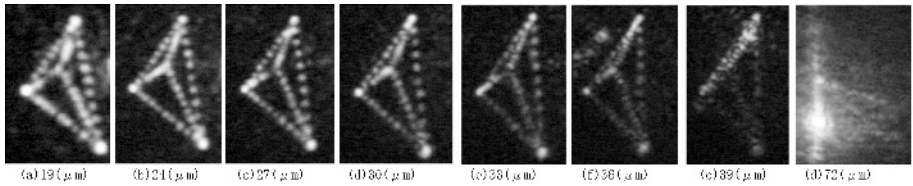


Fig. 8. Example of reconstructed images

3.4 Example of Reconstructed 3D Visual Images of Moving Objects Performed in Our Transmitting System

In the previous section, we showed that in order to transmit and reconstruct nice 3D visual images of moving objects in the holography, encoding rate should be more than 700kbps (PSNR=16dB). Here we present some examples of 3D visual images performed in our transmitting system in Fig.9. They are produced from the CGH file of moving objects transmitted over the network with the encoding rate of 1000kbps, and are obtained by photographed with a constant time interval. Judging from the samples in Fig.9, relatively nice 3D visual images of moving objects are to be reconstructed in our system. But when the network makes unstable operation, the variation in the encoding rate is not to be ignored and multiple reconstructed holographic images are often found in the display. In addition, rarely, 3D images are displayed as the ones of non-moving objects in a certain time interval. Remark that a frame dropping occurs in some unstable network when the encoding rate of 1500kbps is selected. We have shown a graph of the amount of information in the communication (bucket information: kByte) at each time in a day (Fig.10). Each of the data is recorded as a mean value in five minutes. Frame dropping (in Fig.10) seems to have occurred on account of heavy duty of communication, and the amount of bucket information seems to distribute widely. (see Fig.11) From the above results, we can conclude that the best encoding rate is just around 1000kbps in our transmitting system. This seems to be in fact best encoding rate to perform a holographic reconstruction of 3D visual images of moving objects in the network transmission. The de-

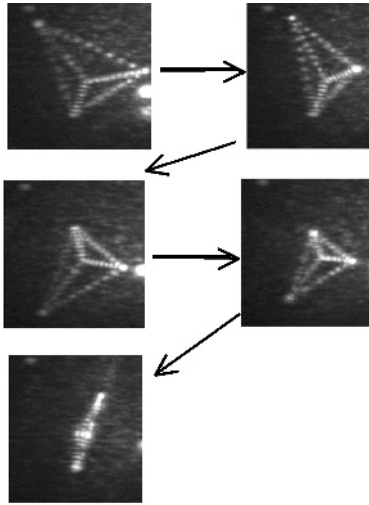


Fig. 9. Samples of reconstructed holographic 3D visual images in the system

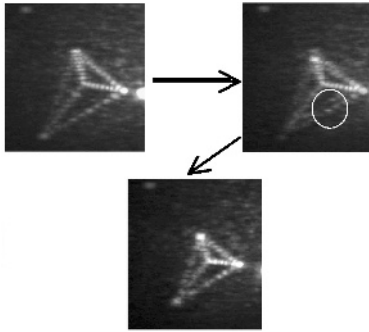


Fig. 10. Sample of 3D holographic image when CGH data dropped the frame

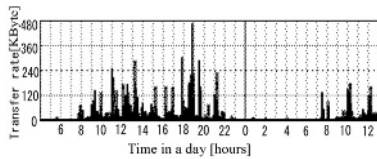


Fig. 11. Amount of bucket information vs. time in a day

tailed relation between frame dropping and amount of bucket information in the communication should be investigated more precisely, however, we would like to discuss it more deeply in another paper.

4 Conclusion

In this paper, an elementary transmitting system of holographic 3D visual images of moving objects is studied applying network streaming process. Here an experiment of on-demand transmission of 3D visual images of moving objects is made over the network using a hologram of Real Media form produced by the computer. From the results, the transmission of holographic 3D visual images of moving objects is found to be possible if PSNR of CGH is around 14dB, and an excellent reconstruction of visual images is found to be performed in high contrast with the background noise to be suppressed minimally. Further, the encoding rate of around 1000kbps is considered to be most suitable. In the later, we would like to investigate the effectiveness of the application to internet system.

References

1. N. Wakamiya, M. Murata, H. Miyahara: Cooperative video streaming mechanisms with video quality adjustment. *APSITT*. **2001** (2001) 106–110
2. K. Shimada, R. Wakabayashi, H. Suzuki, K. Muto, et al.: A Distributing Experiment of Forum with Two Satellite Communication Systems Connected and an Evolution of Picture. *IEICE(Japan)*. **J81-B-II.5** (1998) 486–495
3. K. Takano, K. Sato: Color Electro-Holography by Virtual Image Reconstruction Using Single DMD Panel. *IEICE(Japan)*. **J86-D-II.6** (2003) 869–876
4. A.N. Netravali, B.G. Haskell: *Digital Pictures:Representation, Compression, and Standards* (2nd Ed), (1995) Plenum Press, New York

Coping with Unreliable Peers for Hybrid Peer-to-Peer Media Streaming

Sung-Hoon Sohn, Yun-Cheol Baek, and Juno Chang

Division of Computer Software
Sangmyung University
Seoul 110-743, Korea
{shson, ybaek, jchang}@smu.ac.kr

Abstract. In this paper, we propose a hybrid peer-to-peer streaming architecture for large scale media streaming service. The proposed architecture combines peer-to-peer system and centralized server to exploit advantages of the two models. Then, we deal with streaming load allocation problem under the proposed scheme. Given a set of centralized servers and a set of unreliable supplying peers with heterogeneous bandwidth offers, we show how to assign streaming load to a subset of supplying peers, while considering unreliable performance of each peers. Our experimental results show that the proposed scheme, compared with legacy server with similar capacity, increases the number of clients about 67% on the average.

1 Introduction

Traditional client-server architecture along with voluminous nature of multimedia files makes it difficult to provide large scalable multimedia streaming service. A powerful centralized server with a high-bandwidth Internet connection is somewhat easy to deploy, it has a lot of shortcomings in scalability, reliability, high cost, and load on backbone networks. To deal with these problems, architectures such as proxy caching [7,10] or content distribution network (CDN) [1, 8] are introduced. These architectures employ intermediate nodes called proxy server (in caching) or edge server (in CDN) to duplicate some of multimedia contents to geographically closer nodes to clients. However, they still suffer from the same problems as the client-server case, since they are fundamentally based on client-server paradigm from the viewpoint of proxy server or edge server.

Recently peer-to-peer streaming system has gained a lot of attention as an alternative to existing streaming service architectures. Compared with client-server based models, peer-to-peer streaming system is more scalable in terms of the number of concurrent users and provides much larger streaming capacity in a cost-effective manner. However, pure peer-to-peer media streaming service in real world is almost infeasible due to the following non-technical reasons. (1) Peers are very autonomous entities; they join and leave the network whenever they want to (even in the middle of streaming), since they do not care about overall system's serviceability, availability, etc. (2) Peers are extremely unreliable; they

may suffer from performance degradation due to network congestion, etc. (3) It is very difficult to service unpopular contents by pure peer-to-peer system only, since peers have inclination to access and store popular contents only. Therefore, in order to deploy peer-to-peer paradigm in media streaming service in real world, there must be a scheme to make up for such weak points in the pure peer-to-peer system.

There have been many works dealing with pure peer-to-peer media streaming systems [2,4,5,6,9,12]. Especially, several multi-source on-demand media streaming systems have been proposed. [14] addresses the media data assignment problem in non-parallel fashion and fast system capacity amplification method for multi-source media streaming. Similar to our work, a hybrid system which integrates peer-to-peer into CDN is proposed in [13]. Different from our work, peers are regarded as reliable entities. The authors assume that peers are always up, have no bandwidth degradation, and never stop streaming anyway. Moreover, once a media file is dispersed throughout the system, subsequent streaming requests for that media file are served by peers without intervention of centralized server. They call it handoff and try to optimal handoff time for a given media file. After handoff time, the hybrid system is regressed to pure peer-to-peer system. Even in the middle of hybrid period, a media streaming session is serviced either by CDN server only or by peers only.

Modeling of peers in previous works is too ideal. As mentioned above, peers are never reliable in real world, and it does not make sense to provide deterministic service guarantee with unreliable peers. In this paper, we first propose a hybrid peer-to-peer streaming architecture for large scale media streaming service. The proposed architecture combines peer-to-peer system and centralized server to exploit advantages of the both models. Then, we tackle a problem of streaming load allocation under the proposed architecture. Given a set of centralized servers and a set of peers with heterogeneous bandwidth offers, we suggest a policy to select subset of available centralized servers and supplying peers for streaming a media file. Especially we take account of the unreliable property of peers.

Our contribution can be divided into two parts. First, we propose a hybrid peer-to-peer media streaming architecture integrated with centralized client-server one. Sec-ond, we solve a streaming load allocation problem under the proposed architecture to cope with unreliable peers while maximizing the number of concurrent users. We conduct comprehensive performance evaluation of the proposed scheme. Our results show that that the proposed scheme, compared with legacy server with similar capacity, increases the number of clients about 67% on the average.

The rest of this paper is organized as follows. In Section 2, we propose a hybrid peer-to-peer streaming scheme and explain the system operations under the proposed architecture. In Section 3, we identify the streaming load allocation problem based on the model and propose a load allocation scheme and its admission control algorithm which continues streaming service against some peer failure or degradation. We describe the simulation setup and discuss the results

of performance evaluation in Section 4. Finally, we present our conclusions in Section 5.

2 System Operations

In this section, we present a hybrid peer-to-peer media streaming architecture and explain overall system operations based on the architecture.

Fig. 1 shows the proposed hybrid peer-to-peer streaming architecture. The hybrid system consists of a few streaming servers, an index server, and a set of ordinary peers. Major roles of centralized servers in our architecture are: (1) legacy streaming servers that participate in hybrid streaming session as a server, (2) source of all media files in the system, i.e. a seed peer in the peer-to-peer network. Hereafter, we call them *source peers* in the sense of "source of all media files." Compared with ordinary peer, source peer is a true server in the meaning that they are always up and in operation. The number of source peers in the system is dependent on the scale of the network and client population.

An *index server* of peer-to-peer network knows all the information about which peer is dead or alive, which peer owns which media files, which peers are joining which streaming sessions, and so on. Each *peer* is either a supplying peer for a requesting peer depending on the role of the peer in the streaming session. A peer may be both supplying peer and requesting peer at the same time. Before receiving any streaming service, the client is a requesting peer. After finishing streaming service, the requesting peer caches the media file in disk and it becomes a supplying peer of the media file. A supplying peer may participate in several streaming session as a server at the same time. The heterogeneity of supplying peers is modeled by its maximum number of out-bound sessions, upper limit on the aggregate out-bound bandwidth, and the degree of reliability. We assume that each peer has enough disk storage to contain several video files.

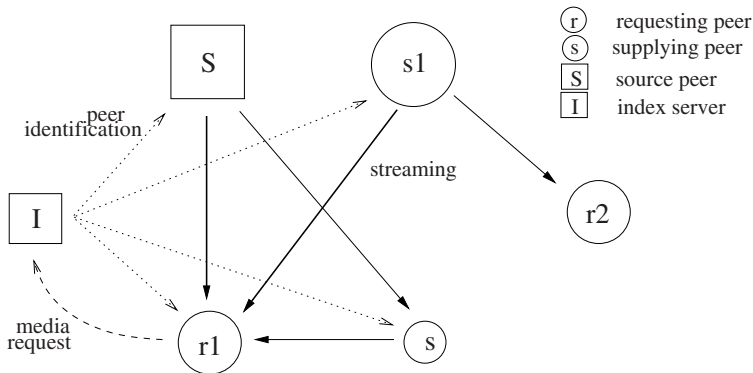


Fig. 1. A hybrid peer-to-peer streaming

The peer-to-peer media streaming in the proposed architecture operates as follows. When a media file is first introduced to the system, it is copied to each source peer¹. At the beginning, there are no supplying peers. Only source peers stream the media data to receiving peers. After a streaming session, the requesting peer reports to the index servers that it has become a sending peer with *contribution parameters* such as a limited out-bound bandwidth contributed to the system's streaming capacity and a limited number of streaming sessions it would support. The values of the two parameters are dependent on the capacity of the peer such as processor power, network bandwidth, and storage.

A peer-to-peer streaming session involves at least one source server and several supplying peers. In order to guarantee continuous playback of media file, the sum of their out-bound bandwidth contribution is at least the same as the media playback rate. On receipt of media playback request, the index server first determines source peers in the neighborhood of the requesting peer. It also checks if there are active supplying peers who own the media file such that (1) they have an available out-bound slot and (2) the sum of their out-bound bandwidth, including source peers, is greater than or equal to the playback rate of the media. If so, the request will be served by the selected supplying peers and source peers servers; otherwise, the request will be served by the source peers only.

For a given media file and possible candidate supplying peers (including centralized servers), there may exist many possible ways to select a subset of candidate supplying peers who will participate in streaming session of the media file. There are certain criteria in designing peer selection policy. These include (1) guaranteeing continuous playback, (2) to minimize initial buffering delay, (3) to be resilient from (sending) peer degradation or failure, (4) to maximize the remaining streaming capacity of the system, (5) to minimize load on network, (5) to disseminate the media file as fast as possible. The principal objective of the streaming service is to guarantee continuous playback. In legacy client-server paradigm, continuous playback is guaranteed by reliable centralized server. However, in peer-to-peer paradigm, peers acting as a server are not reliable at all. They may go down, suffer from degradation or failure. Moreover, one cannot predict these failures in advance. Therefore, the unreliable property of peers makes it difficult to guarantee continuous playback. In what follows, we propose a peer selection policy considering unreliable property of peers.

3 Streaming Load Allocation

In this section, we try to answer the following question; *how to distribute a streaming load among source peers and ordinary supplying peers?* We first define clearly the streaming load allocation problem, and then we suggest a streaming load allocation policy in consideration of peer's reliability. We also propose

¹ A media file can be published by an ordinary peer. In this case the media file should be duplicated to source peers before being serviced to peers

a failure resilience scheme dealing with degradation of peers in the middle of streaming session.

Before discussing the streaming load allocation, we first model the heterogeneity of peers as follows. Each peer has two attributes concerning heterogeneity; the degree of reliability and out-bound bandwidth limit. Let the *degree of reliability* of each supplying peer i be specified as a percentage p_i of the total amount of media data that is supposed to arrive on time. That is, in worst case, a requesting peer receives only p_i of the data the supplying peer i tries to send. If a peer is a source peer, p_i is always 1; otherwise, for ordinary supplying peers, it is less than 1 (p_i is even zero when p_i is down). Let the *out-bound bandwidth limit* of peer i be specified as f_i^{max} . Namely, peer i limits its out-bound bandwidth contribution by f_i^{max} .

Now consider a media streaming session of m possible candidate peers (including source peers). We assume that supplying peers can service the requesting peer by proceeding in periodic rounds, retrieving and sending a fixed amount of media data for each round. Let f_1, f_2, \dots, f_n denote the amount of media data sent in each round. The problem of streaming load allocation is to find n ($n \leq m$) and f_i ($i = 1, \dots, n$), which satisfy the following inequality:

$$p_1 * f_1 + p_2 * f_2 + \dots + p_n * f_n \geq q * F \quad (1)$$

subject to $0 \leq p_i \leq 1$, $0 < f_i \leq f_i^{max}$, and $\sum_{i=1}^n f_i^{max} \geq F$, where F is the total amount of data needed by requesting peer during a round for best-quality playback and q is the streaming quality requirement provided by requesting peer. The left hand side of the equation represents the lower bound on the expected amount of media data received during a round in worst case. The right hand side means the amount of media data that is needed by client while satisfying the client-supplied QoS parameter².

This problem can be solved using the following algorithm. First of all, we should include one or more source peers, if available, in order to guarantee minimum quality of streaming. Then, given m candidate peers, we sort the supplying peers according to their values of $p_i * f_i^{max}$ in decreasing order. We start with the largest value peers and assign its f_i to be $\alpha * p_i * f_i^{max}$, where α is an appropriate constant less than 1. We then continue to assign to each supplying peer this value, beginning with the ones with larger values and moving to the ones with smaller values, until the above equation is satisfied.

Now we discuss the failure resiliency feature of the proposed scheme. By failure resiliency, we mean a recovery from the situation where peer's degree of reliability is changing due to some reasons such as network congestion on links between supplying peer and receiving peer, abrupt overload on supplying peers, etc. Consider a simple case involving two supplying peers. For supplying peer s_1 , p_1 is 1.0 and f_1 is 10 and, for supplying peer s_2 , p_2 is 0.6 and f_2 is 10. Then, the amount of media data received by requesting peer per each round is $1 * 10 + 0.6 * 10 = 16$. After a while, the degree of reliability of s_1 changes to 0.9,

² The equation can be used as an admission control criteria by index server when admitting a new requesting peer

which may results in the decrease of streaming quality in the near future. We can avoid the situation by simply changing the values f_i of other supplying peers who shows stable degree of reliability. In our example, we increase the value of f_2 by 2, we can easily maintain the streaming quality such as $0.9*10+0.6*12 \geq 16$.

4 Experiments

We evaluate the performance of the proposed architecture through extensive simulation experiments. We present the simulation setup and the results in this section.

We use the Network Simulator ns-2 [11] in the experiments. We use large Internet-like network topology generated by GT-ITM topology generator [3], add peers via DSL or LAN to the routers in the topology. We use 112 media files of 30-minute durations each recorded at a rate of 192Kb/s. Each peer has 128Kb/s out-bound bandwidth of 2-6 concurrent streaming sessions.

We simulate the following scenario. First source peers introduce media files into the network. According to the uniform arrival pattern, a peer joins the network and requests a media file. Media files are selected according to the zipfian distribution with skew factor of 0.7. Then, the streaming steps described in Section 2 and 3 are put into operation. If the request can be satisfied, i.e., there is a sufficient capacity in the system, connections are established between the supplying peers (and source peers) and the requesting peer and the streaming session begins. The send and receive over UDP and carries CBR traffic. When the streaming session is over, the requesting peer caches the whole media file.

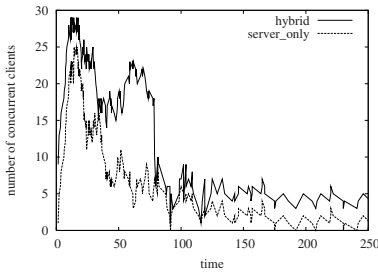


Fig. 2. Number of concurrent users

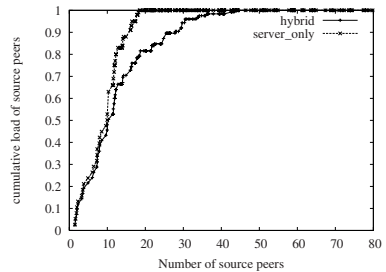


Fig. 3. Load of source peers

Fig. 2 and 3 shows the performance of the proposed hybrid streaming scheme compared with the legacy CDN service. In this experiment, 600 clients make 4,100 requests during 270 minute simulated interval. For legacy streaming service, we use 10 streaming servers and, for hybrid peer-to-peer service, we use 3 source peers. As shown in Fig. 2, the proposed scheme accepts much larger number of client's requests. More specifically, the number of concurrent clients increases by 67.2% on average. Fig. 3 shows the cumulative load on the source

peers in hybrid system and on the streaming servers in legacy system. Source peers in the hybrid system are much less loaded due to the help of ordinary supplying peers.

In order to find the optimal number of source peers in the hybrid system streaming service, we measure the effects of the number of source peers on the reject ratio of client's request with various source peer capacities. As shown in Fig. 4, when the number of source peers is relatively small, the reject ratio decrease drastically as the number of source peers increases. However, more source peers does not affect greatly when the number of source peers are large.

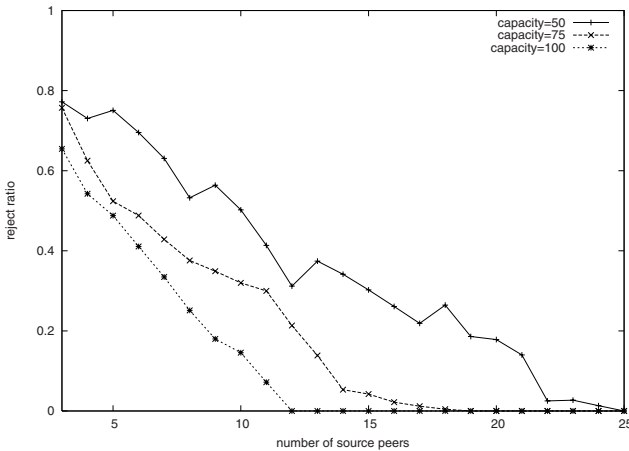


Fig. 4. Effects of number of source peers on reject ratio

5 Conclusions

In this paper, we propose a hybrid peer-to-peer media streaming architecture integrated with legacy client-server architecture. We solve the streaming load allocation problem under the proposed scheme to maximize the number of concurrent users for a given set of peers. The proposed hybrid architecture and the load allocation scheme has the following features: (1) the proposed scheme takes advantage of both client-server streaming and pure peer-to-peer streaming, (2) at least one source peer is involved with each streaming session, which results in more stable streaming quality, (3) by considering the degree of reliability of peers in streaming load allocation, the scheme can be easily applicable to real-world service, (4) the allocation scheme easily adapts to peer's failure or network congestion against unreliable peers. We conduct comprehensive performance evaluation of the proposed scheme. Our results show that the proposed scheme, compared with legacy server with similar capacity, increases the number of clients about 67% on the average.

References

1. Akamai. <http://www.akamai.com>.
2. C-star. <http://www.centerspan.com>.
3. Calvert, K., Doar, M., and Zegura, E.: Modeling Internet Topology. *IEEE Transactions on Communications*, pages 160–163, December (1997)
4. Castro, M., Druschel, P., Kermarrec, A., Nandi, A., Rowstron, A., and Singh, A.: SplitStream: High-bandwidth Content Distribution in a Cooperative Environment. In *Proceedings of International Workshop on Peer-to-Peer Systems* (2003) 000–000
5. Deshpande, D., Bawa, M., and Garcia-Molina, H.: Streaming Live Media over a Peer-to-Peer Network. Stanford Database Group Technical Report **2001-20** (2001)
6. Gummadi, K.P., Dunn, R.J., Saroiu, D., Gribble, D., Levy, H.M., and Zahorjan, J.: Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload. In *Proceedings of ACM Symposium on Operating Systems Principles* (2003) 000–000
7. Jin, S., Bestavros, A., and Iyengar, A.: Accelerating Internet streaming media delivery using network-aware partial caching. In *Proceedings of IEEE ICDCS 02*, Vienna, Austria, July (2002)
8. Nguyen, T. and Zakhor, A.: Distributed Video Streaming over Internet. In *Proceedings of SPIE/ACM MMCN* (2003) 000–000
9. Padmanabhan, V.N., Wang, H.J., Chou, P.A., and Sripanijkulchai, K.: Distributing Streaming Media Content Using Cooperative Networking. In *Proceedings of NOSSDAV* (2002) 000–000
10. Sen, S., Rexford, J., and Towsley, D.: Proxy prefix caching for multimedia streams. In *Proceedings of IEEE INFOCOM 99*, New york, USA, (1999)
11. The Network Simulator – ns2.: <http://www.isi.edu/nsnam/ns>
12. Tran, D.A., Hua, K.A., and Do, T.T.: A Peer-to-Peer Architecture for Media Streaming. *IEEE Journal on Selected Areas in Communications*, **22(1)** (2004) 000–000
13. Xu, D., Chai, H., and Kulkarni, S.: Analysis of a Hybrid Architecture for Cost-Effective Media Distribution. In *Proceedings of SPIE/ACM Conference on Multimedia Computing and Networking*, Santa Clara, CA (2003)
14. Xu, D., Hefeeda, M., Hambruch, S., and Bhargava, B.: On Peer-to-Peer Media Streaming. In *Proceedings of ICDCS* (2002) 000–000

Evaluation of Token Bucket Parameters for VBR MPEG Video

Sang-Hyun Park¹ and Yoon Kim²

¹ Department of Multimedia Engineering, Sunchon National University,
Sunchon, 540-742, Korea
Tel: +82-61-750-3833, Fax: +82-61-750-3590
shark@sunchon.ac.kr

² Kangwon National University, Chunchon, Korea
yooni@ieee.org

Abstract. Guarantees of quality-of-service (QoS) in the real-time transmission of video is a challenging task for the success of many video applications. To guarantee the QoS, it is necessary to provide traffic sources with the capability of calculating the traffic characteristics to be declared to the network on the basis of a limited set of parameters statistically characterizing the traffic and the required level of QoS. In this paper, we develop an algorithm for the evaluation of the traffic parameters which characterize the video stream when a QoS requirement is given. To this end an analytical traffic model for the VBR MPEG video is introduced. Simulation results show that the proposed method can evaluate the traffic parameters precisely and efficiently.

Keywords: Video Traffic Modeling, Token Bucket, Admission Control.

1 Introduction

The transmission of the VBR compressed video requires an accurate traffic modeling to meet the QoS of the video and manage the network resources efficiently. When the video calls request to enter the network, the network admission controller decides if it can provide the calls with their negotiated QoS without violating the QoS guarantees of existing calls in the network. Accurate traffic modeling and analysis of the QoS parameters enable the network admission controller to make decisions which ensure the integrity of the traffic sources and the efficiency of the network [1,2].

The network elements need to be informed of the traffic characteristics of a real-time application to reserve network resources for the application. The IETF has specified the token bucket algorithm as the traffic policing mechanism in the IntServ and the Diff-Serv. That is, setting up a flow over the Internet requires the traffic specification for the flow to be specified in advance in terms of the token bucket traffic parameters which include the token generation rate, the token bucket size, and the peak rate [1,2,3].

To determine the proper traffic parameters, accurate traffic modeling and its queueing analysis are required. In this paper, we propose a new activity-based traffic model for the VBR MPEG video to analyze a token bucket traffic shaper at the network access point. Using the proposed traffic model, we propose an analytical method to calculate the loss probability that packets do not comply with the specifications of the token-bucket traffic shaper.

This paper is organized as follows. In Section 2, we propose a traffic model for VBR MPEG video. In Section 3, we develop an analytic method to analyze the parameters of the token bucket traffic shaper. In Section 4, we illustrate the experimental results of the proposed methods. Finally, the conclusion remarks are presented in Section 5.

2 Traffic Model for VBR MPEG Video

In the proposed traffic model for the VBR MPEG video, we first construct the GOP layer model to represent the video activity and then the frame layer model is developed to represent the periodic pattern of the MPEG video.

Let $x^G(n)$ denote the number of bits obtained by coding the n th GOP. Depending on the amount of activity in the actual scene, the sequence $x^G(n)$ has different statistical properties during different time intervals. Let $x^G(i; j)$ be the GOP sequence of $\{x^G(i), x^G(i+1), \dots, x^G(j)\}$ which are associated with the same scene. If $x^G(i; j)$ and $x^G(k; l)$, respectively, are associated with low and high activity scenes, they must have different statistical characteristics. In both sequences, the sizes of two successive GOP's (e.g., $x^G(i)$ and $x^G(i+1)$) are highly correlated because of the continuity in the actual video scene. This leads to use a DAR process as the model [4]. However, the correlation between the sizes of successive GOP's and the mean and variance of the number of bits per GOP are not the same for $x^G(i, j)$ and $x^G(k, l)$. Thus, we use a DAR process with time-varying parameters as the model.

According to the discussion above, we classify the video sequence into L activity bands according to the size of each GOP. That is, the GOP whose size is between the predetermined thresholds γ_{j-1} and γ_j , belongs to the j th activity band. Here, γ_{j-1} and γ_j are selected in such a way that a first-order AR process sufficiently models the temporal behavior of the j th activity band [5]. The number of activity bands L depends on the nature of the video scenes. Then, the video activity can be modeled by a Markov chain with L activity states which represent L activity bands. The transition matrix $P_{band} = [p_{i,j}]$ of the Markov chain can be estimated as follows:

$$p_{i,j} = \frac{\text{number of transitions } i \text{ to } j}{\text{number of transitions out of } i}. \quad (1)$$

To model GOP's associated with the i th activity band, we use a DAR(1) process [4]. Let $\{\epsilon_i(n)\}$ and $\{v_i(n)\}$ be two sequences of i.i.d random variables. The random variable $\epsilon_i(n)$ has a discrete state space and distribution h_i . The random variable $v_i(n)$ is a Bernoulli random variable with $P\{v_i(n) = 1\} =$

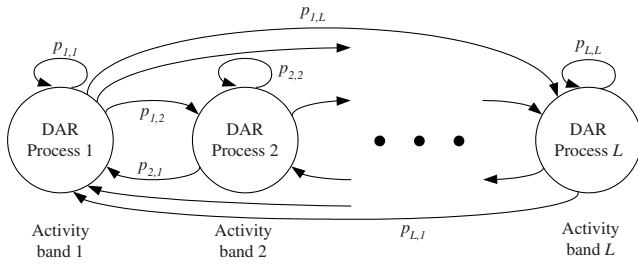


Fig. 1. Activity-based MPEG video traffic model at GOP layer

$1 - P\{v_i(n) = 0\} = a_i$. For the DAR(1) process, a_i is the correlation coefficient between the sizes of two successive GOP's. A GOP sequence for the i th activity band is modeled by the DAR(1) process as follows:

$$x^G(n) = v_i(n)x^G(n - 1) + \{1 - v_i(n)\}\epsilon_i(n). \tag{2}$$

Here, to generate $\{\epsilon_i(n)\}$, we use the arrival rate histogram of GOP's associated with the i th activity band. To compute a histogram, we break up the range of values covered by the data set into B disjoint intervals $[d_0^i, d_1^i), [d_1^i, d_2^i), \dots, [d_{B-1}^i, d_B^i)$ where the interval width $\Delta d_i = d_j^i - d_{j-1}^i$ is constant. Here, d_0^i and d_B^i , respectively, are set to the minimum size and the maximum size of GOP's associated with the i th activity band. Let z_j^i be the average size of the samples that lie in the interval $[d_{j-1}^i, d_j^i)$, and $\{\omega(n)\}$ be an i.i.d. sequence with the uniform distribution $U(0, 1)$. Given GOP sizes, z_j^i ($j = 1, \dots, B$) and the frequency h_j^i which is the proportion of the samples falling into the interval $[d_{j-1}^i, d_j^i)$, the sequence $\{\epsilon_i(n)\}$ is generated by

$$\epsilon_i(n) = z_j^i \text{ with } j = \min \left\{ l : \sum_{j=1}^l h_j^i > \omega(n) \right\}. \tag{3}$$

The overall traffic model of the GOP layer is shown in Fig. 1. The proposed model produces a GOP sequence using the following DAR(1) process whose parameters are determined by the state of the Markov chain. Let $\hat{x}^G(n)$ and b_n , respectively, denote the number of bits and the state of Markov chain of the n th GOP generated by the traffic model. According to the proposed model, $\hat{x}^G(n)$ is obtained using the size of the previous GOP and states b_{n-1} and b_n , i.e.,

$$\hat{x}^G(n) = \begin{cases} v_i(n)\hat{x}^G(n - 1) + \{1 - v_i(n)\}\epsilon_i(n), & \text{if } b_n = b_{n-1} = i \\ \epsilon_i(n), & \text{if } b_n \neq b_{n-1} \text{ and } b_n = i \end{cases} \tag{4}$$

In (4), we assume that the numbers of bits in the GOP's before and after the activity state transition are independent. That is, the number of bits in the first

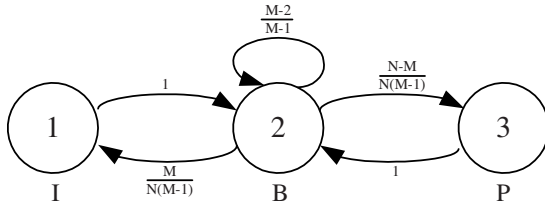


Fig. 2. Frame-level Markov chain

GOP following the activity state change is generated randomly according to the arrival rate histogram. Then, a_i is easily obtained by

$$a_i = 1 - \frac{c_i}{2 \cdot \text{var}(x^G(n)|b_n = i)} \tag{5}$$

with

$$c_i = E[\{x^G(n) - x^G(n - 1)\}^2 | b_n = b_{n-1} = i]. \tag{6}$$

In each GOP, I-, B-, and P-frames appear periodically according to the pre-defined GOP parameters. The GOP structure is specified by two parameters (N, M) , where N is the distance between two successive I-frame and M is the distance between I- and subsequent P-frame or two successive P-frames. In the proposed model, we use a frame-level Markov chain to generate a frame sequence according to the pre-defined GOP parameters. The transition of frame types can be described as a probability model as in Fig. 2. If we let $\hat{x}^T(n)$ be the number of bits of frame type T in the n th GOP, $\hat{x}^T(n)$ is determined by

$$\hat{x}^T(n) = \frac{\mu^T}{\sum_{k \in \{I, B, P\}} \mu_k N_k} \hat{x}^G(n), \quad T \in \{I, B, P\}$$

where μ_T is the mean value of the number of bits of frame type T and N_T is the number of frames of type T in a GOP.

3 Evaluation of Traffic Shaper Parameters

In this section, we present a queuing analysis method of the proposed traffic model for the VBR MPEG video. We assume that the data generated from one frame are evenly spaced over one frame duration.

For each frame, states of the queuing system employed are defined in terms of the quantities (s_n, t_n, b_n) , where s_n , t_n , and b_n , respectively, are the DAR process state of, the frame type state of, and the activity state of the n th frame. Note that the stochastic process $\{s_n, t_n, b_n\}$ is a Markov chain with state space $\{(i, j, k) : 1 \leq i \leq B, 1 \leq j \leq 3, 1 \leq k \leq L\}$.

Let $p(i, j, k, i', j', k')$ be the probability that the Markov chain $\{s_n, t_n, b_n\}$ moves from the (i, j, k) state to the (i', j', k') state. When the probability $p(i, j, k, i', j', k')$ is mapped into 2-dimensional space through the mapping,

$$l = i + B(j - 1) + 3B(k - 1) \quad \text{and} \quad l' = i' + B(j' - 1) + 3B(k' - 1),$$

the resulting transition matrix, P , which is of order $3BL$ with elements $p(l, l')$, takes the form

$$P = \begin{bmatrix} A_1 & B_{1,2} & \cdots & B_{1,L} \\ B_{2,1} & A_2 & \cdots & B_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ B_{L,1} & B_{L,2} & \cdots & A_L \end{bmatrix}, \quad (7)$$

where

$$A_i = \begin{bmatrix} 0 & I & 0 \\ \frac{M}{N(M-1)}p_{i,i}C_i & \frac{M-2}{M-1}p_{i,i}I & \frac{N-M}{N(M-1)}p_{i,i}I \\ 0 & I & 0 \end{bmatrix} \quad (8)$$

and

$$B_{i,j} = \begin{bmatrix} 0 & 0 & 0 \\ p_{i,j}\hat{C}_j & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (9)$$

Here, C_i represents the i th activity band and is given by $a_i I - (1 - a_i)\hat{C}_i$ and \hat{C}_i is a matrix whose rows are the arrival rate histogram of the i th activity band, $(h_1^i, h_2^i, \dots, h_B^i)$. Also, 0 and I , respectively, are the $B \times B$ zero and identity matrices. In addition, the bit rate matrix Λ representing the bit rate at state (i, j, k) is given by $\text{diag}(\lambda_{1,1,1}, \lambda_{2,1,1}, \dots, \lambda_{B-1,3,L}, \lambda_{B,3,L})$ where $\lambda_{i,j,k}$ denotes the bit rate of the (i, j, k) state.

Next, we approximate the state transition matrix P which is a discrete time Markov chain with a continuous time Markov chain $Q = \tau(P - I)$ where τ is the number of frames per second of the VBR MPEG video source and I is the $3BL \times 3BL$ identity matrix.

To analyze a token bucket traffic shaper, we use a fluid approach [6]. In [6], the virtual buffer W is defined as $X - Y + K_T$ where X and Y , respectively, represent the occupancies of the smoothing buffer and the token bucket and K_T is the token bucket size.

Let $W(t)$ be the content of the virtual buffer, $S(t)$ be the DAR process state, $T(t)$ be the frame type state, and $B(t)$ be the activity state at time t . Let $F_{i,j,k}(t, x)$, $t \geq 0$, $x \geq 0$, be the probability that at given time t , the DAR process state is i , the frame type state is j , the activity state is k , and the buffer content does not exceed x as

$$F_{i,j,k}(t, x) = \Pr\{W(t) \leq x, S(t) = i, T(t) = j, B(t) = k\} \quad (10)$$

$$(i \in \{1, 2, \dots, B\}, j \in \{1, 2, 3\}, k \in \{1, 2, \dots, L\}).$$

The forward transition equation from time t to $t + \Delta t$ is written as

$$\begin{aligned}
 & F_{i,j,k}(t + \Delta t, x) - F_{i,j,k}\{t, x - (\lambda_{i,j,k} - R)\Delta t\} \\
 &= \sum_{i'=1}^B \sum_{j'=1}^3 \sum_{k'=1}^L q_{i',j',k'}^{i,j,k} F_{i',j',k'}(t, x)\Delta t + O(\Delta t),
 \end{aligned} \tag{11}$$

where $q_{i',j',k'}^{i,j,k}$ is the $\{i' + B(j' - 1) + 3B(k' - 1), i + B(j - 1) + 3B(k - 1)\}$ -th element of Q and R is the token generation rate. In the steady state, it is

$$(\lambda_{i,j,k} - R) \frac{\partial}{\partial x} F_{i,j,k}(x) = \sum_{i'=1}^B \sum_{j'=1}^3 \sum_{k'=1}^L q_{i',j',k'}^{i,j,k} F_{i',j',k'}(x). \tag{12}$$

Let us define the $F(x)$ and D matrices as

$$F(x) = [F_{1,1,1}(x), F_{2,1,1}(x), \dots, F_{B-1,3,L}(x), F_{B,3,L}(x)] \tag{13}$$

$$\text{and } D = \Lambda - RI. \tag{14}$$

Rewriting (12) in matrix form gives

$$\frac{\partial}{\partial x} F(x)D = F(x)Q. \tag{15}$$

It is well-known by [7] that the solution of (15) is given by

$$F(x) = \sum_{l: \text{Re}[z_l] < 0} \alpha_l \phi_l e^{z_l x} + \pi \tag{16}$$

where π is the stationary probability vector of Q ; $\{z_l, \phi_l\}$ is the eigenvalue and eigenvector pair of the eigenvalue problem of $z_l \phi_l D = \phi_l Q$.

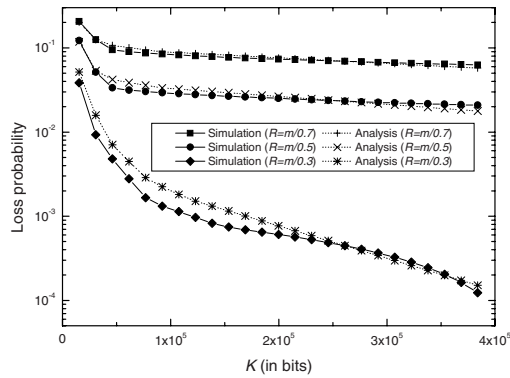
Since every bit requires a token for transmission, the average token throughput must equal the average throughput. Thus, the average throughput λ^* is given by

$$\lambda^* = R - \sum_{i=1}^B \sum_{j=1}^3 \sum_{k=1}^L (R - \lambda_{i,j,k}) F_{i,j,k}(0). \tag{17}$$

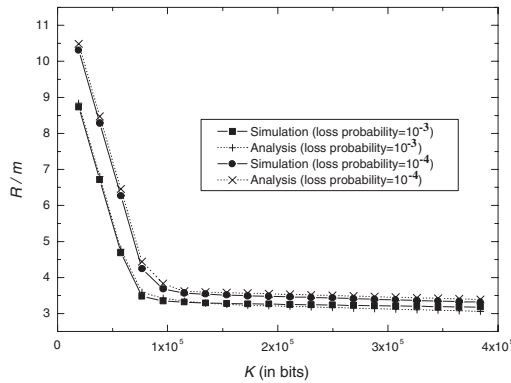
Finally, the loss probability is given by

$$P_L = 1 - \frac{\lambda^*}{\sum_{i=1}^B \sum_{j=1}^3 \sum_{k=1}^L \lambda_{i,j,k} \pi_{i,j,k}}. \tag{18}$$

The smoothing buffer size K_D is determined by the peak rate p and the maximum acceptable delay [2,3]. For a given loss probability, all possible pairs of $\{R, K_D + K_T\}$ can be analytically derived using (18). Then, the token bucket size K_T is also determined.



(a)



(b)

Fig. 3. Comparison of simulation and analytic method: (a) loss probability and (b) token generation rate

4 Experimental Results and Discussion

Simulation and analysis results are presented to illustrate the accuracy of the proposed model. For our experiment, we use the *Star Wars* MPEG video sequence. The GOP parameters (N, M) of the video sequences are $(12, 3)$ and the number of frames per second is 30 frames/s.

We first compare the simulation results obtained using the real video sequences with the proposed analysis results obtained using (18). Fig. 3(a) shows the loss probability of the *Star Wars* when the token generation rate R is $m/0.3$, $m/0.5$, and $m/0.7$, where m is the mean bit rate and K is $K_D + K_T$. It is seen that the results of the proposed method are very similar to simulation results. We also investigate the token generation rate which satisfies the loss requirement for the given value of K . Fig. 3(b) shows the token generation rate divided by the mean bit rate of the *Star Wars* as a function of K when the maximum

acceptable loss probability is 10^{-3} and 10^{-4} . It can be observed that the results of the proposed analysis method and the simulation results match quite well in both regions.

5 Conclusion

In this paper, we have proposed a new activity-based MPEG video traffic model by considering the overall variation of the video activity. Based on the proposed traffic model, we have presented the analysis method which can determine the loss probability of the VBR MPEG video. Using the proposed analysis method, we can determine the token bucket parameters and the output link rate when the acceptable maximum loss probability is given. Experimental results show that the proposed traffic model and analysis method accurately estimate the performance of the VBR MPEG video traffic.

The proposed traffic model and analysis method can be effectively used for the transmission of the VBR MPEG video applications over networks, such as video-on-demand and video conferencing.

References

1. M. A. El-Gendy, A. Bose, and K. G. Shin, "Evolution of the Internet QoS and Support for Soft Real-Time Applications," *Proc. IEEE*, vol. 91, pp. 1086–1104, July 2003.
2. A. Lombardo, G. Schembra, and G. Morabito, "Traffic Specifications for the Transmission of Stored MPEG Video on the Internet," *IEEE Transactions on Multimedia*, vol. 3, pp. 5–17, Mar. 2001.
3. M. F. Alam, M. Atiquzzaman, and M. A. Karim, "Traffic shaping for MPEG video transmission over the next generation internet," *Computer Communications*, vol. 23, pp. 1336–1348, 2000.
4. D. Heyman and T. V. Lakshman, "Source Models for VBR Broadcast-Video Traffic," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 40–48, Feb. 1996.
5. N. D. Doulamis, A. D. Doulamis, G. E. Konstantoulakis, and G. I. Stassinopoulos, "Efficient Modeling of VBR MPEG-1 coded Video Sources," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 93–112, Feb. 2000.
6. M. Schwartz, *Broadband Integrated Networks*. New Jersey: Prentice Hall PTR, 1996.
7. D. Anick, D. Mitra, and M. M. Dondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell System Technical Journal*, vol. 61, pp. 1871–1894, Aug. 1982.

Perceptual Video Streaming by Adaptive Spatial-Temporal Scalability*

Wei Lai¹, Xiao-Dong Gu², Ren-Hua Wang¹, Li-Rong Dai¹, and
Hong-Jiang Zhang²

¹ Department of Electronics Engineering and Information Science, University of
Science and Technology of China, Hefei, Anhui, 230027, China

laiwei@ustc.edu, {rhw, lrdai}@ustc.edu.cn

² Microsoft Research Asia, 5F, Sigma Building, Zhichun Road, Beijing, 100080, China
{t-xiaogu, hjzhang}@microsoft.com

Abstract. To maximize user satisfaction of video streaming, there is a tradeoff between spatial quality (image clarity) and temporal quality (motion smoothness) under a limited bandwidth. How to balance the requirement for the two aspects is a subjective selection. In this paper, we firstly introduce how to efficiently measure the subjective perceived spatial and temporal quality. Then we present a visual perception model to predict viewer's satisfaction given the perceived spatial quality and temporal quality based on the result of a user study. At last, an adaptive video streaming system utilizing the perception model is proposed, which can automatically choose dropping frames or cutting the scalable bitstream according to the variable bandwidth, and obtain maximized perceptual quality.

1 Introduction

Network visual communication has become an active research area in recent years. One of the most challenging problems for the implementation of a video communication system is that the available bandwidth of the networks is usually insufficient for the delivery of the voluminous amount of the video data. To stream video over the dynamic and unpredictable best-effort Internet, it is essential that the sending rate of the video is adaptive to achieve the best quality. Recently, several techniques have emerged that are very promising and may lead to significantly improved video codecs in comparison with the current standards.

The first technique is to incorporate Human Visual System (HVS) models into coding system. It is well accepted that perceived video quality does not correlate well with PSNR. HVS characteristics must be considered to provide better visual quality measurements.

The second one is to develop continuous rate scalable coding algorithms. The FGS (Fine Granularity Scalable) technique [1] is one of the best coding techniques that make it possible to take every available bit to enhance video

* This work was performed at Microsoft Research Asia.

spatial quality (PSNR) under variable available bandwidth. Integrating with the HVS characteristics, the perceived spatial quality can be maximized for any given bandwidth.

Dropping frames is another major technique for rate adaptation [2]. Frame dropping will cause motion judder and decrease the perceived temporal quality (motion smoothness) especially for video sequences with a high motion. The objective of the frame rate adapting technique is to optimize the temporal quality given a bandwidth.

Both spatial quality (image quality, clarity) and temporal quality (motion quality, smoothness) are required to achieve a satisfactory viewing experience of video. However, under a given bitrate, especially a low bit rate, a tradeoff between spatial quality and temporal quality is necessary to maximize user's satisfaction. The rules of this tradeoff should be determined by human's sensitivities to image clarity and temporal smoothness, and the effects on spatial and temporal quality by PSNR decreasing and frame dropping given a bandwidth.

In this paper, we present a perception model to describe an integral perceived quality given the spatial quality and temporal quality, and establish a perception priority function based on it. Then, we propose an adaptive video streaming system, which will choose dropping frame (decrease temporal quality) or cutting FGS layer bitstream (decrease spatial quality) automatically, to maximize the perceived quality along with the changing of bandwidth by time.

The remaining of this paper is organized as follows. Perceived spatial quality and perceived temporal quality are introduced in section 2 and section 3 respectively. Section 4 gives the perception model learned from the user study. In section 5, the perception model is utilized in a video streaming system. Section 6 gives a simulation of the system and section 7 concludes the paper.

2 Perceived Spatial Quality

Presently, the objective quality measure Peak Signal-to-Noise Ratio (PSNR) is widely employed to evaluate video quality. However, it is well accepted that perceived video quality does not correlate well with PSNR. Human Visual System (HVS) characteristics must be considered to provide better visual quality measurements. Foveation based HVS model is one of the practical HVS model. It is based on the feature that the spatial resolution of the HVS is the highest around the point of fixation (foveation point) and decreases rapidly with increasing eccentricity [3].

The foveated distortion (FD) of a decoded digital image from the original image is defined as:

$$FD = \left(\frac{1}{M} \sum_{n=1}^M (S_f(v, f, \mathbf{x}_n) \cdot |p(\mathbf{x}_n) - p'(\mathbf{x}_n)|)^2 \right)^{\frac{1}{2}} \quad (1)$$

where $S_f(v, f, \mathbf{x}_n)$ is defined in [3], M is the number of pixels of the digital image, and $p(\mathbf{x}_n)$ and $p'(\mathbf{x}_n)$ are the corresponding pixel values of the original image and the target image.

And we define a perceived spatial quality PSQ as (c is a constant):

$$PSQ = c/FD \quad (2)$$

To get better subjective perception, foveated image can be generated by emphasizing the regions near the foveation points and deemphasizing the rest regions. The foveation points can be generated automatically by saliency-based visual attention model [4], or set by manual mark. Based on the MPEG 4 codec, the bit allocation within a frame is weighted by $S_f(v, f, \mathbf{x})$ in [3]. The bit density R (bit/pixel) of pixel \mathbf{x} is set as (R_0 is reference bit density):

$$R(\mathbf{x}) = R_0 * S_f(v, f, \mathbf{x}) \quad (3)$$

3 Perceived Temporal Quality

Dropping frames is one of the major techniques for rate adaptation in adaptive video streaming without spatial scalability. Dropping frames will cause motion judder since the dropped frames usually are replaced by replaying previous frames. In fact, a viewer's perceived motion judder and his/her satisfaction to frame dropping heavily depends on the motion in video sequence. It is found that frame repetition gives good results where no motion is present, but fails in moving areas, resulting in clearly visible motion judder in frame rate up-conversion applications [5]. Dropping frame with camera pan motion is more annoying than dropping frames with other kinds of motion as motion judder is most noticeable on camera pans in video [6].

To model user satisfaction to frame dropping, the motion description feature should embody such characteristics of human perception. In our work, a low-level feature named SWM (segmentation weighted motion magnitude) is introduced to describe the motion in video sequence. The SWM is defined as the weighted sum of motion magnitudes of the dominant motion segmentations. For each frame, the MVs (motion vectors) of the MVF (motion vector field) are clustered by the similarity and coherence of the direction, magnitude and position of the MVs, and the MVF is segmented into several regions.

Suppose the MVF of frame t is segmented to N regions, R_1, R_2, \dots, R_N . The size (number of MVs/MBs) of R_i is S_i ($i = 1, 2, \dots, N$). The mean motion magnitude of R_i is M_i . The SWM is defined as ($ThrS$ and $ThrM$ are thresholds):

$$SWM = \frac{\sum_i S_i \cdot M_i}{\sum_i S_i} \quad \text{where } S_i > ThrS \text{ and } M_i > ThrM; \quad (4)$$

The larger the SWM is, the more motion judder will be perceived by viewers when frames are dropped, and the poorer visual quality is supposed to be. If some frames are dropped, for example, the even frames are dropped to get a 1/2 frame rate video sequence. It is equal to the magnitudes of the MVs (M_i) are doubled. So the perceived temporal quality PTQ is defined as (λ is a constant):

$$PTQ = \exp\left(-\frac{SWM}{\text{framerate} \cdot \lambda}\right) \quad (5)$$

4 Modeling Visual Perception

To model the priority of perceived spatial quality (image clarity) and perceived temporal quality (motion smoothness) of video presentation, we conduct a user study to evaluate the overall perceptual quality under different spatial quality and temporal quality.

In our user study tool, two presentations of a same video shot with different frame rate and bits allocated pre frame will be shown together. Viewer is required to point out which presentation perceived better to his/her opinion. 150 video shots (each about 5 seconds long) are selected from about 30 movies. Each video shot will be presented in $framerate \in \{4, 6, 8, 12, 24\}$ fps and $bits \in \{4\%, 6\%, 8\%, 12\%, 16\%, 24\%\}$ of a reference bits allocation per frame, and under the combination $framerate * bits \in \{6 * 4\%, 6 * 8\%, 6 * 16\%, 12 * 16\%\}$. In each presentation the video shot is shown in $framerate$ (fps), each frame is foveated as (3) and use bits percentage of the reference bits allocation. The perception comparison will only be conducted in different presentations for same video shots: first between two presentations in same $framerate * bits$, then the worst perceived presentation in a higher level $framerate * bits$ will be compared with the best perceived presentation in a lower level one. For each ($framerate, bits$) presentation of every video shots, the perceived spatial quality PSQ and perceived temporal quality PTQ will be computed.

5 viewers are invited to manually label each of the couples (PSQ_1, PTQ_1) , (PSQ_2, PTQ_2) by comparison.

“>”: The first presentation perceived better;

“<”: The second presentation perceived better;

“=”: No obvious difference between the two presentations.

It is necessary to point out that the labeling process is independent among the viewers. All of the viewers are students of non-computer science specialty. All viewers are required about 1 meter away from the screen.

At last, we use the comparison labeling result to generate a Perception Priority function: $PP(PSQ, PTQ)$. For two video presentations, one has the perceived spatial and temporal quality (PSQ_1, PTQ_1) and another has (PSQ_2, PTQ_2) . If the first one gets a better perceived quality than the second one, we can get $PP(PSQ_1, PTQ_1) > PP(PSQ_2, PTQ_2)$, vice versa.

5 Perception Priority Based Video Streaming System

Our Streaming system is based on FGS (fine granularity scalable). FGS provides the continuous scalability of the distortion-rate performance. The FGS video-coding technique is described in the Amendment of MPEG-4 [1].

Based on the perception priority results, we establish a priority based video delivery model. When the bandwidth is insufficient for transmit full quality and frame rate video stream, the delivery system will make a choice between dropping frames (decrease PTQ) or cutting FGS layer of the stream (decrease PSQ) based on the perception priorities of corresponding video presentation (PSQ, PTQ) .

Suppose the compressed video stream has a fixed GOP structure “IBBPBBPBBPIBBP” and we call a 3-frame-group like “IBB” or “PBB” a sub-GOP. In our priority based delivery system, I and P frames are given higher priorities because of decoding interdependency. For a continuous interval of frames, the first B frame of a sub-GOP is assigned lower priority than the second B frame, these B frames are dropped to get a 2/3 frame rate. Afterwards, the second B frames of the sub-GOPs are dropped to get a 1/3 frame rate.

The video sequence is pre-processed: foveated frame is made for each frame as (3). In the FGS layer, the bit planes of emphasized regions are shifted up to make sure that the emphasized regions get a higher enhancement priority during the video streaming process. When the video sequence is compressed, the reference PSQ and reference PTQ of each frame t are computed at the same time, namely $RSQ(t)$ and $RTQ(t)$ respectively. $RSQ(t)$ is the PSQ of foveated frame t under a reference bit density $RefR$ (bit/pixel). $RTQ(t)$ is the PTQ of frame t at full frame rate.

The adaptive streaming process is applied on every short interval, which contains K sub-GOPs ($3K$ frames, $2K$ B frames and K I/P frames). According to the bandwidth of that moment (frame t , the beginning of the interval), we can get the available bits for this interval, defined as $bits(t) = bandwidth(t) * 3K / framerate$. And we suppose the ratio of the bits number of the K I/P frames to the bits number of the $2K$ B frames in this interval is r_PB . Thus, we will get 3 delivery schemes for this interval:

(1). Full frame rate. No frames are dropped. The FGS layer is cut off according the bandwidth. The bits for the $2K$ B frames is about $bits(t)/(r_PB + 1)$, and the left bits are for the I/P frames (Fig.1 (a)).

(2). 2/3 frame rate. The first B frames of the sub-GOPs are dropped. The bits for the left K frames is about $bits(t)/(r_PB + 0.5)$, and the left bits are for the I/P frames (Fig.1 (b)).

(3). 1/3 frame rate. All the B frames are dropped, and all the $bits(t)$ is assigned to I/P frames. The FGS layer is cut off to accord with the bits number (Fig.1 (c)).

According to the Rate-Distortion model [7]: $R(D) = \gamma \ln(\frac{1}{\alpha D}) \implies D(R) = \frac{1}{\alpha} \exp(-\frac{R}{\gamma})$, here D represents the distortion like FD , and by (2), we have: $PSQ \propto \exp(R/\gamma)$. And we can estimate the PSQ of frame t by $RSQ(t)$ given the bit density R (bit/pixel): $PSQ(t, R) \approx RSQ(t) \cdot \exp(R/\gamma) / \exp(RefR/\gamma)$.

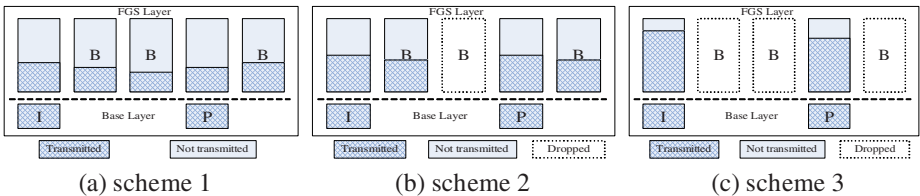


Fig. 1. Bitstream of the 3 delivery schemes

We can get the PTQ of a sub-GOP at frame t by $RTQ(t)$:

(i) if frame $t + 1$ and frame $t + 2$ is not dropped: $PTQ(t, 1) = RTQ(t)$;

(ii) if frame $t + 1$ is dropped, it is equal to $1/2$ frame rate at frame t , according to (5): $PTQ(t, 1/2) = (RTQ(t))^2$;

(iii) if frame $t + 1$ and frame $t + 2$ are all dropped, it is equal to $1/3$ frame rate at frame t : $PTQ(t, 1/3) = (RTQ(t))^3$.

Then, we can get the mean PSQ and mean PTQ of the above 3 delivery schemes:

Scheme 1:

$$\overline{PSQ_1(t)} = \frac{1}{3K} \sum_{n=0}^{K-1} (PSQ(t + 3n, R_{IP}) + PSQ(t + 3n + 1, R_B) + PSQ(t + 3n + 2, R_B)) \tag{6}$$

where: $R_{IP} = \frac{r_{PB}}{r_{PB}+1} bits(t) \frac{1}{K * framesize}$, $R_B = \frac{1}{r_{PB}+1} bits(t) \frac{1}{2K * framesize}$ and

$$\overline{PTQ_1(t)} = \frac{1}{3K} \sum_{n=0}^{3K-1} (PTQ(t + n, 1)) \tag{7}$$

Scheme 2:

$$\overline{PSQ_2(t)} = \frac{1}{2K} \sum_{n=0}^{K-1} (PSQ(t + 3n, R_{IP}) + PSQ(t + 3n + 2, R_B)) \tag{8}$$

where: $R_{IP} = \frac{r_{PB}}{r_{PB}+0.5} bits(t) \frac{1}{K * framesize}$, $R_B = \frac{1}{r_{PB}+0.5} bits(t) \frac{1}{K * framesize}$ and

$$\overline{PTQ_2(t)} = \frac{1}{2K} \sum_{n=0}^{K-1} (PTQ(t + 3n, 1/2) + PTQ(t + 3n + 2, 1)) \tag{9}$$

Scheme 3:

$$\overline{PSQ_3(t)} = \frac{1}{K} \sum_{n=0}^{K-1} (PSQ(t + 3n, R_{IP})) \tag{10}$$

where: $R_{IP} = bits(t) \frac{1}{K * framesize}$ and

$$\overline{PTQ_3(t)} = \frac{1}{K} \sum_{n=0}^{K-1} (PTQ(t + 3n, 1/3)) \tag{11}$$

Then, we simply choose the scheme which has the maximum perception priority:

$$\text{the choosen scheme number} = \arg \max_{i \in \{1,2,3\}} \{PP(\overline{PSQ_i(t)}, \overline{PTQ_i(t)})\} \tag{12}$$

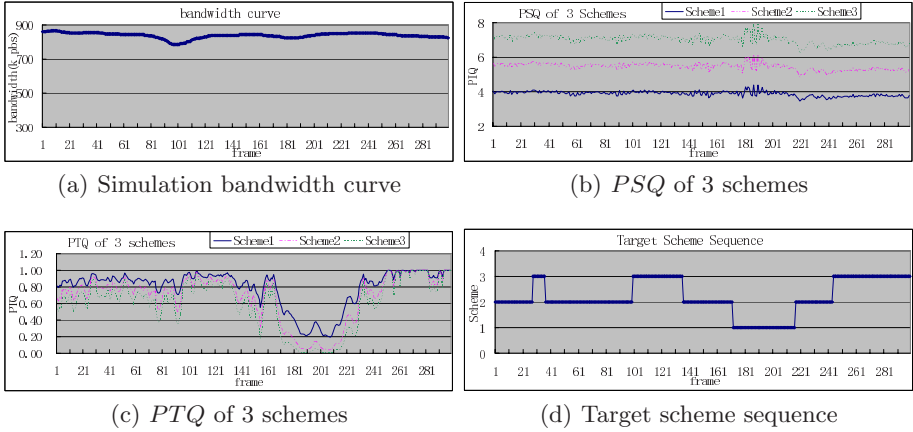


Fig. 2. Simulation results

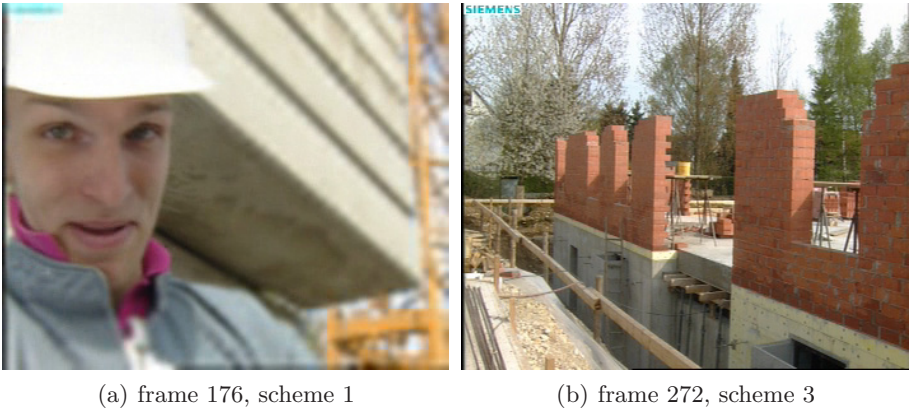


Fig. 3. Streaming results

6 Simulations

As an example, a simulation is applied on the test sequence *foreman*. The simulation bandwidth curve of 300 frames is shown in Fig.2(a). After all frames are made foveated frame, the *PSQ* of each frame of 3 delivery schemes are estimated, shown in Fig.2(b). The *PTQ* of each frame of the schemes are also computed, shown in Fig.2(c).

The target delivery scheme of each time interval ($K=3$ sub-GOPs, 9 frames) is decided by (12), and is shown in Fig.2(d).

Fig.3 shows 2 frames of the delivered video sequence by the target scheme sequence. Fig.3(a) is frame 176, due to the high motion of the interval, cutting

FGS layer to decrease PSQ has a higher priority than dropping frames to decrease PTQ , so scheme 1 is chosen (full frame rate to maintain smooth motion, but the spatial quality is decreased by allocate less bits to the background region) to get the best perceptual quality of this interval. While Fig.3(b) is frame 272, the motion is slight, all the B frames can be dropped without introduce evident motion judder, then scheme 3 (1/3 frame rate, high spatial quality) is chosen to get the best perceptual quality with a high PSQ of this interval.

By adaptively choose scheme under the changing available bandwidth during the streaming process, the delivered video sequence can get best perceived visual quality.

7 Conclusions

In this paper, we have presented the measurement of subjective perception on spatial quality (image clarity) and temporal quality (motion smoothness). A perception model is learned from user study to predict the subjective visual perception of a video presentation with a spatial quality and a temporal quality. By this model, an adaptive video streaming system is proposed to maximize the visual quality of the delivered video stream by automatically choose dropping frames (decrease temporal quality) or cutting the scalable bitstream (decrease spatial quality) along with the variable bandwidth. The simulation results have proved the effectivity of the streaming system.

References

- [1] W.P.Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard", IEEE Trans. On Circuits and Systems for Video Technology, Vol.11, No.3, pp. 301-317, Mar. 2001.
- [2] T.M.Liu, H.J.Zhang and F.H.Qi, "Perceptual Frame Dropping in Adaptive Video Streaming", Proc. of ISCAS 2002, Arizona, the US, May 2002.
- [3] Z. Wang, A. C. Bovik, L. Lu and J. Kouloheris, "Foveated wavelet image quality index", Proc. SPIE, Application of digital image processing XXIV, vol. 4472, pp. 42-52, July-Aug. 2001.
- [4] Itti L, Koch C. "A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems". SPIE Human Vision and Electronic Imaging IV (HVEI'99), Vol. 3644, pp. 373-382, San Jose, CA, January 1999.
- [5] R. Castagno, et al, "A Method for Motion Adaptive Frame Rate Up-conversion", IEEE Trans. On Circuits and Systems for Video Technology, Vol. 6, pp. 436-446, Oct. 1996.
- [6] <http://www.microsoft.com/hwdev/TVBROADCAST/TempRate.htm>
- [7] Tihao Chiang and Ya-Qin Zhang, "A New Rate Control Scheme Using Quadratic Rate Distortion Model", IEEE Trans. on Circuits and Systems for Video Technology, Vol.7, No.1, pp. 246-250, Feb.1997.

A Proxy Caching System Based on Multimedia Streaming Service over the Internet

Hui Guo, Zhou Jingli, Zeng Dong, and Yu Shengsheng

Department of Computer Science and Engineering,
Huazhong University of Science and Technology, Wuhan 430074, China,
{hguo, jlzhou, dzeng, ssyu}@wtwh.com.cn

Abstract. As real-time streaming media is becoming a significant proportion of network traffic, interest in exploring caching system for streaming media objects has increased. However, existing techniques for caching Web objects are not appropriate for the multimedia streaming service. In this paper, we focus on the proxy caching problem specifically for multimedia streaming objects. A prototype design and implementation of a proxy caching system - HUSTProxy is proposed. The main contribution of HUSTProxy is its ability of partial video caching, prefix caching, and sending rate control. These techniques are implemented in our system and are detailed discussed in this paper. By validate our implementation, the experimental results show that the partial video caching can contribute to economize disk cache space and the rate control mechanism can contribute to reduce service startup latency as expected.

1 Introduction

Most recent research on proxy caching focus on handles generic web objects, such as Harvest [1] and Squid [2]. However, the existing techniques for caching Web objects are not appropriate for the continuous streaming media services because video files are usually much larger than web documents. Compared to Web caching techniques, evaluation of proxy caches for multimedia streaming objects are still immature, design and evaluation of multimedia proxy caching mechanisms clearly require substantially more investigation.

This paper presents the design and implementation of a multimedia proxy caching system which contribute to reduce server load, network load and service start-up latency, named as HUSTProxy. Different from earlier works, we are emphasize on partial caching and partial replacement techniques. In particular, we propose that proxy caches only a fix set of frames at the beginning of each popular stream, and the replacement scheme is from the end portions of an unpopular stream. Combining with prefetch techniques implemented in our design, the proxy caching disk space can be dramatically decreased. Another contribution of server-proxy-client architecture is its ability to adjust sending rate to client based on the available bandwidth between the proxy and interested clients path. Our study complements this efforts to perform rate control, the effectiveness of this proposal is validated through experimental results from

prototype implementation. As expected, by increasing the data sending rate to requested clients, the service start-up latency can be improved accordingly. Additionally, we propose a new definition of popularity of a cached stream compared with earlier work in this paper.

The rest of this paper is organized as follows. The next section describes the design and architecture of our proxy caching system. Section 3 validates our implementation through various experiments, and presents the preliminary results. Section 4 gives a briefly reviews of related work. In the last section, we present the conclusions and directions for future work.

2 Design and Implementation of HUSTProxy

2.1 System Architecture

As the HUSTProxy is located in the network, it appears as a client for a server and as a server for a client. Figure 1 depicts the internal architecture of HUSTProxy as a combination of a client and a server. Showing as Fig.1, the video streaming was controlled through RTSP/TCP [3] sessions. There were two sets of sessions for the client. The first was established between the originating video server and proxy to retrieve uncached blocks. The other was between the proxy and client. Each of video and audio was transferred over a dedicated RTP/UDP [4] session. The quality of service was monitored over RTCP(RTP Control Protocol)/UDP [5] sessions. We will explain the functionality of individual components by describing Request Manage Module within the HUSTProxy in section 2.2.

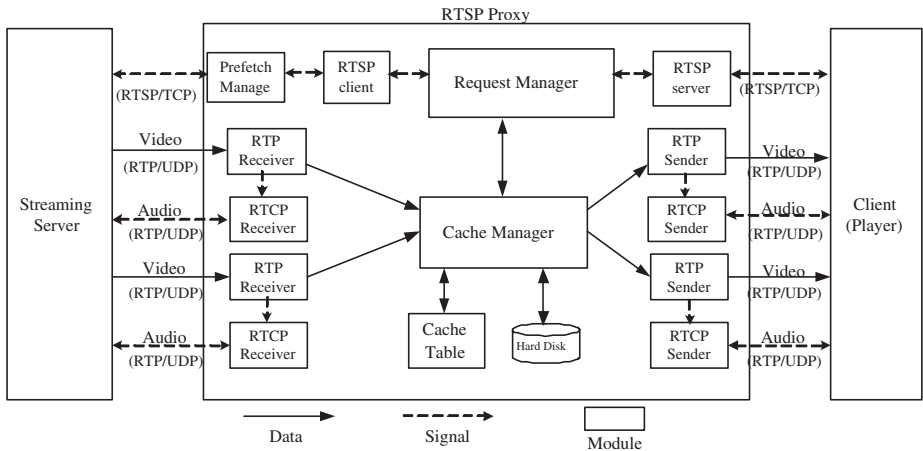


Fig. 1. Proxy system architecture

2.2 Request Management

Request Manager module handles all the RTSP signaling between HUSTProxy and client or server. A complete process of streaming proxy deal with a client RTSP request described as follows. First, a client begins by establishing connections for audio/video streams with the proxy server using a series of RTSP OPTIONS, DESCRIBE, and SETUP messages. These RTSP messages are received by the Request Manager through an RTSP Server module. The proxy server relays RTSP OPTIONS, DESCRIBE, and SETUP messages to the video server. Thus, connections between the video server and the proxy server are also established. Then, the client requests delivery of the video stream by sending an RTSP PLAY message. On receiving a PLAY request for a video stream from a client, the Request Manager begins providing the client with data blocks. It first examines the cache table. As long as a copy of requested stream is available in the cache, the Request Manager sequentially reads them out and sends them to the client through the RTP Sender. Simultaneously, The Request Manager forwards the PLAY request with a modified Range header, which is to prefetch the remainder of the requested stream in advance for later streaming response.

If the requested stream s is missed, the Request Manager simply relay the RTSP PLAY message to the video server with no modification. On receiving a stream from the video server through the RTP Receiver, the Request Manager sets its flag with on to indicate that the streaming blocks is being transmitted, and it relays the blocks to the RTP Sender. When reception is completed, the flag is cancelled and the Request Manager deposits the blocks in its local cache disk. If there is not enough room to store the newly retrieved block, the Cache Manager replaces one or more less important segments in the cache with the new stream segment. When a proxy server receives an RTSP TEARDOWN message from a client, the proxy server relays the message to the video server, and closes the sessions.

2.3 Packets Recombination

When the user requests a video in the cache, it is served by sending to them the portion of the video locally present in cache, while obtaining the remainder frames from the streaming server and transparently passing it on to the client. In order to achieve prefix caching, video servers and streaming protocols must support random access, and the proxy caching system must incorporate prefetch technique. For implementing prefetching function, we could use the RTSP PLAY message with a modified Range header forward to server, which based on how much data is available from the local disk, e.g., Range: 20-End. The number 20 means that the last cached time of prefix caching.

Since the RTP packets obtained from a particular source begin with a random base timestamp and a random base sequence number, the proxy have to compose a coherent stream from the local disk source and the server source. Figure 2 shows an example of how to compose RTP streams from two different sources into a consistent one. Here, we consider a media clip which is 15 seconds long; the initial

5 seconds are cached in proxy while the rest of the 10 seconds data is obtained from a remote server. And assuming the stream clockrate is 90000, which can be obtained from SDP feedback from the server. Through the timestamp of the last RTP packet, base timestamp and the stream clockrate, we can obtain the last cached time with the metric of seconds by equality (1)

$$CachedTime(seconds) = \frac{LastTimestamp - BaseTimestamp}{Clockrate} \quad (1)$$

Showing as the Figure 2, 180000 and 630000 are the timestamps of the first and last RTP packets obtained from the disk; ω is the timestamp of the next packet that has not been cached on the disk, the timestamp gap between two successive RTP packets is 6000. Assume -376000 and -16000 are the timestamps of the first and last RTP packets obtained from the network. Before forwarding the first packet obtained from the network, the HUSTProxy needs to determine the outgoing sequence number and timestamp for this packet. In Fig.2, the RTP packets of segment 1 are obtained from playing of the beginning of the stream, the random base timestamp of this stream should be 180000. We can calculate that the cached time position in the disk is $(630000 - 180000)/90000 = 5(seconds)$. And the value of ω should be $(630000 - 180000) + 6000 = 456000$. Whereas the actual value we obtained from the network is -376000. So the packets gap between the two sessions is $G = 456000 - (-376000) = 832000$. By the session gap, we can obtain that the outgoing timestamp of the first packet from the network is $-376000 + G = -376000 + 832000 = 456000$, and timestamp of the last packet from the network is $-16000 + G = -16000 + 832000 = 816000$. The SSRC field in the RTP headers fetched from network can be set the same as value of RTP headers which stored in cache disks.

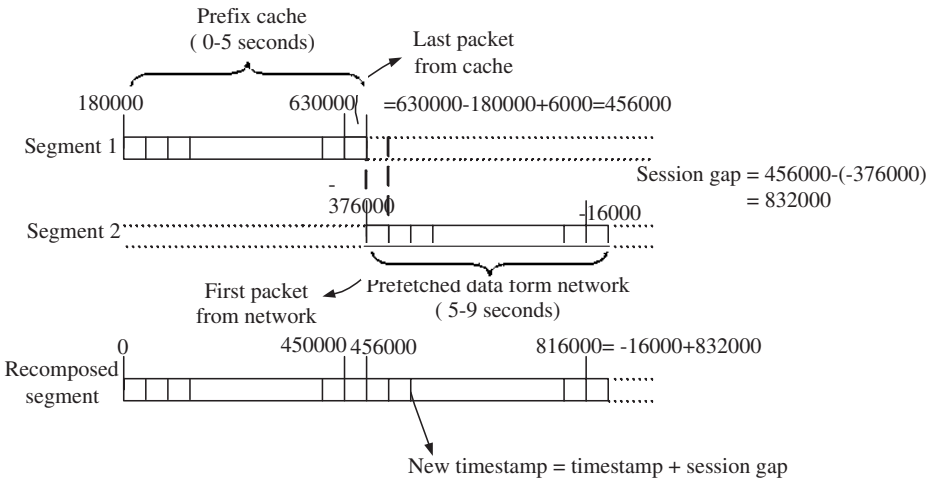


Fig. 2. Packets recombination

3 Performance Experiments

3.1 Network Load Reduction

The goal of a proxy is to intercept client request transparently and reduce the number of requests sent to the server, which led to reduction of server load and network load. To demonstrate the decrease of network load by our HUSTProxy system, we use traffic variation ratio (TVR) as the criterion of evaluation.

$$TVR = \frac{\xi_{sp}(t)}{\xi_{pc}(t)} \tag{2}$$

$\xi_{sp}(t)$:The total amount of data transferred from the server to proxy at time t .
 $\xi_{pc}(t)$:The total amount of data transferred from the proxy to client at time t .

A larger TVR value indicates increase in relative network load. On the contrary, a smaller TVR value means a decrease in network load. The maximum value of TVR is 1. In our experiments, we use a server-proxy-client setup, where the client simulates multiple media players by our RTSP request generation program. This program disregards the response data, only generating requests for media objects according to Zipf-like distribution model. Simultaneously, we detect the network traffic both of the server-proxy path and proxy-client path with the increase of time t .

Figure 3 shows that as time increasing, the traffic variation ratio variation with different prefix caching size. The server has 200 MPEG media objects which have the same playback length of 60 seconds. In each time, we specify a fixed prefix caching size, which is 20%, 40%, 60% and 80% of the total cached streams

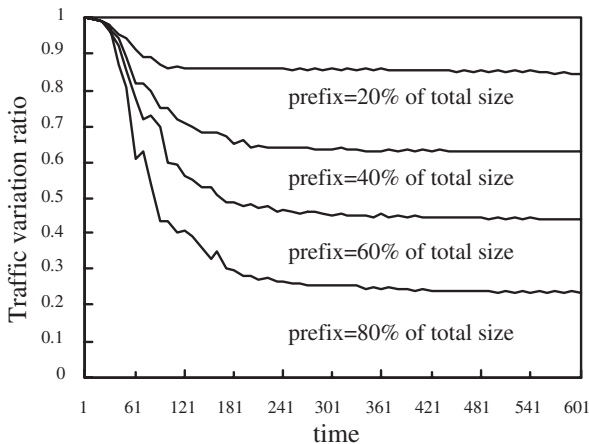


Fig. 3. Traffic variation ratio with variation of prefix caching size

Table 1. Parameters and definition

Parameter	Definition
t_1	The delay between the server and the proxy
t_2	The delay between the proxy and the client
S	Start playback buffer length (the scale is seconds)
S_1	The first S_1 seconds of the stream has been cached
α	Data transfer rate from server to client
β	Data transfer rate from proxy to client

long respectively. We can see that the *TVR* changes more rapidly with the increase of prefix caching size. Larger of *TVR* variation indicates larger of network load decrease.

3.2 Start-Up Latency Reduction

Another major function of a proxy system is to reduce the client request start-up latency. When a client receives stream data from a server or the proxy, it does not start playing the object until its playout buffer is filled to absorb network jitter. So there is a start-up latency at the beginning of playback every time. Now we consider two scenarios: one involves a server-client architecture and the other is server-proxy-client architecture. For presenting the client start-up latency expression, we define the related parameters as table 1.

We assume the delay between the server and the proxy is t_1 , and the proxy to client path delay is t_2 . Then the delay between the server and the client is $t_1 + t_2$, and the round-trip delay is $2t_1 + 2t_2$. Considering server-client architecture, the media server sends out data packets according to its playback rate α bytes/second, and assume each client keeps a playout buffer of S seconds, the initial S seconds of data are related to client's start-up latency. The start-up latency of this scenario T_0 can be given by $T_0 = 2t_1 + 2t_2 + S$. Next, we consider the server-proxy-client architecture. if $0 \leq S_1 \leq S_2$, the proxy starts two processes concurrently. One is download the existing S_1 seconds of data to the client as β transfer rate, which takes $S_1\alpha/\beta$ seconds to transfer S_1 seconds of data. The other process is to request the rest of $S - S_1$ seconds of data from stream server. It takes $2t_1$ seconds for the first byte of the data to arrive proxy. So the time for both process to finish is $\max((S_1\alpha)/\beta, 2t_1)$. The $S - S_1$ seconds of data is transferred firstly from server to proxy, then from proxy to client. Thus the transfer rate is $\min(\alpha, \beta) = \alpha$. Therefore the transfer time of this part of data is $t_2 + (S - S_1)\alpha/\alpha = t_2 + S - S_1$. We can present the start-up latency of this scenario by equality(3):

$$\begin{aligned}
 T_1 &= t_2 + \max((S_1\alpha)/\beta, 2t_1) + t_2 + (S - S_1)\alpha/\min(\alpha, \beta) \\
 &= 2t_2 + S - S_1 + \max(S_1\alpha/\beta, 2t_1)
 \end{aligned} \tag{3}$$

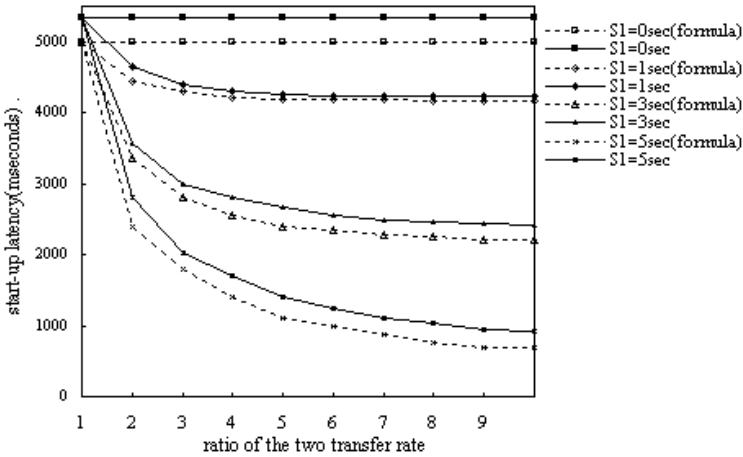


Fig. 4. Client start-up latency with the variation of cached size(seconds)

Figure 4 shows the client start-up latency with the variation of data transfer rate. The solid lines represent experimental results, and the dotted lines represent calculation results from the equality(3). The parameter setting in the formula is based on $t1 = 15ms$, $t2 = 1ms$. In these experiments, the value of β/α is changed from 1 to 10, and the start playback buffer length S is 5 seconds. We set different value of S_1 as 1, 3, 5 seconds respectively for each experiment. From Fig.4, we note that the calculation results and the experimental results appear closely, and with a fixed S_1 , the start-up latency decreases as the increase of β/α . With fixed β/α value, the start-up latency decreases as the increase of S_1 .

4 Related Work

Caching techniques have been widely used for traditional Web content such as HTML pages and image files [1][2][6]. Some early work on the storage for media objects can be found in [7][8]. Sen [9] presents a prefix caching which stores only the first part of the popular media object. Rejaie et al. [10] considered layered-encoded multimedia streams. The proxy attempts to replay a quality-variable cached stream while performing quality adaptation dynamically. Acharya et al. [11] proposed a cooperative caching techniques named as MiddleMan. MiddleMan caches video streams across multiple proxies where they can be replaced at a granularity of a block. However, this work has not implemented fast-forward/rewind play for clients. We have supported this functionality in our HUSTProxy prototype. Some of work have mentioned sending rate control, but they do not give a detailed discussion and experimental results.

5 Conclusions and Future Work

This work describes the implementation a proxy caching system for multimedia streaming service over the Internet. Some issues and challenges in the design of proxy system are detailed discussed, and the solutions are presented. From experimental results, HUSTProxy can reduce overall client start-up latency and reduces network load by intercepting a large number of server accesses.

The implementation of a distribute proxy system is our next target. In this way, all proxies support cooperative each other. On receiving a request, the local proxy responses the request at first. If the cache missed, the proxy forwards the request to other distributed proxies. Issues related to how these proxies cooperate, how to find the most appropriate proxy. And this scheme also needs a global cache management and replacement policy to guarantee cache consistency.

References

1. Chankhunthod A., Danzig P. B., Neerdaels C. A hierarchical Internet object cache. In Proceedings of the USENIX 1996 annual technical conference. January (1996) 153–163
2. Squid web proxy cache. <http://www.squid-cache.org/>
3. Schulzrinne H., Rao A., Lanphier R.: RTSP: Real Time Streaming Protocol, IETF RFC2326, April (1998)
4. Schulzrinne H., Casner S., Frederick R.: RTP: A Transport Protocol for Real-Time Applications, IETF RFC1889, January (1996)
5. Schulzrinne, H.: RTP Profile for Audio and Video Conferences with Minimal Control, IETF RFC 1890, January (1996)
6. Abrams M., Standridge C.R., Abdulla G.: Caching proxies: limitations and potentials. In Proceedings of the WWW conference Darmstadt, Germany. December (1995) 119–133
7. David P.A., Yoshitomo O., Ramesh G.: A File System for Continuous Media. *ACM Transactions on Computer Systems*. **10** (1992) 331–337
8. Tobagi F., Pang J., Baird R.: Streaming RAID: A disk array management system for video files. In Proceedings of ACM Multimedia ACM Multimedia, (1993) 393–400
9. Sen S., Rexford J., and Towsley D.: Proxy prefix caching for multimedia streams. In Proceedings of IEEE INFOCOM. March (1999) 1310–1319
10. Rejaie R.: Multimedia Proxy Caching Mechanism for Quality Adaptive Streaming Applications in the Internet. In Proceedings of IEEE INFOCOM. March (2000) 3–10
11. Acharya S.: Techniques for improving multimedia communication over wide area networks. Ph.D. Thesis. Cornell University. NY. January (1999)

Simulation and Development of Event-Driven Multimedia Session

Nashwa Abdel-Baki and Hans Peter Großmann

University of Ulm, OMI, Albert-Einstein-Allee 43, 89081 Ulm, Germany
nashwa.abdel-baki@e-technik.uni-ulm.de,
hans-peter.grossmann@kiz.uni-ulm.de

Abstract. Networked multimedia systems are recently introduced to the field of research and study. However, the studied and developed systems focus mainly on analysis and testing of the networking protocols and algorithms. There is a real need to consider the user demand with its complexities and diversities. In this paper we demonstrate our system that we developed to fulfil an event-driven multiparty multimedia streaming system. The system enables the user to interact live with the system interface to compose his own session of presentation. According to the developed system, we have considered running a simulation study of a multimedia communication system with different platforms and different network topologies to approach the reality of testing the system over the existing global Internet. The simulation study and the developed system have shown real achievement in the direction of user-driven multimedia sessions.

1 Introduction

Research and development in the field of networking multimedia systems is one of the demanding issues over the Internet Protocol platform. The ongoing activities in this area focus on the networking algorithms to impose the Quality of Service (QoS) architecture on the existing platforms. However, there is a lack of study groups in the direction of user demand. There is a need to dedicate a lot of effort to this area of research and study.

In this paper we are continuing our work on developing a fully integrated interactive multimedia session. We are targeting an event-driven multiparty session. This means the user interactivity is the core of the session construction. This is a challenging process especially in the direction of scalability and system response time.

Scalability mainly is an issue in the direction of networking infrastructure. It is more seriously a considerable issue from the point of view of the number of multimedia sources that the application can handle. Moreover, it is an issue for the system processing power and the operating system and associated multimedia software.

In [1] we introduced our synchronized multimedia prototype system that is based on clock-driven protocol and the dynamic finite state machine (DFSM)

model. The system is capable of handling multiparty multimedia session over a distributed networking platform. Our prototype implementation in [3] has shown promising results, despite the scalability limitations. In this paper we present our current version of the system development, that shows improvement in the direction of scalability and system response time. Parallel to the implementation we present our simulation study of different scenarios of more sophisticated multimedia presentation systems.

The rest of this paper is organized into four sections. Section 2 describes the developed system. Section 3 demonstrates the simulation study with different topologies. We discuss our developed system and the results of the simulation study in section 4. We conclude the work in section 5 as well as our plans for future work.

2 Developed System

In the earlier version of our developed system described in [3] we defined a rendering multimedia system in a distributed environment. The rendering process is realized according to a score file of the access grammar. The core of the system is based on finite state machine model, while the user interface is XML-based implementation.

It was essential for this work to elaborate on the user interactivity mandate, especially in the university environment with the mission of teaching students and knowledge creation and transfer, [2].

Here in this version of our developed system, we offer the functionality of a user driven interface to enable interactively composing the multimedia session on the spot. The user is capable of defining his own sequencing of rendering. He can interact with the system to compose any sequence or group of sequences. The sequencing can be time-based or logic-based. It is a multiparty system mainly from the point of view of multimedia resources, not only in the dimension of the multi-participant architecture.

The system platform enables the user two modes of access. He is capable of accessing the classical precasted form of the presentation. In addition he can formalize his own presentation. This means the user has the ability to control the associated attributes of timing sequences, navigating or skipping between rendered media sources. Consequently, this means the user creates his own event-driven presentation session.

Assuming a multimedia session platform composed of number v of video sequences, a of audio sequences, i of images, and t of text files. A user would like to compose his session in a temporal and logical sequence as shown in Figure 1.

In this session, one of the user wishes in his presentation is to be composed such that Audio1 starts $D1$ minutes after the start of Text1. The user is free to define the time offset between the different media sources of his session. He is also capable of instructing his session to end Video2 and Image7 as a consequence of start of Video5. This means the start of Video5 is the event that drives the end of Video2 and Image7 simultaneously.

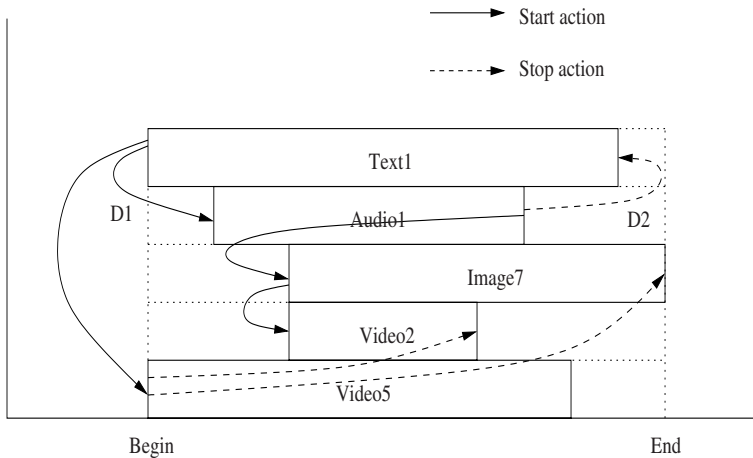


Fig. 1. Event-Driven Multimedia Session

This dynamic user-defined presentation is created interactively so that the user has the chance to examine his own setup and modify or correct it if needed. He is also capable of navigating the archiving media system and selecting the appropriate media sources to render. He can define his presentation session either in a sequential or parallel mode or a combination of both of them.

The developed user interface is based simply on a webserver implementation in addition to using the Hypertext Preprocessor (PHP) programming language to enable the user-defined presentations. We adopted Synchronized Multimedia Integration Language (SMIL) coding to develop the scheduled multimedia sources per session. This current version of our system greatly improves the rendering response time and the overall system performance.

3 Simulation Study

Our system is developed and tested within the university environment. It is essential to measure the performance over the global Internet. As a prerequisite phase of development, we run the test in this stage through a simulation study. In this study we examine streaming sessions over distributed systems. We focus on measuring synchronization and linking between distributed multimedia systems.

In fact modeling and simulating the Internet is not a straight forward process. Our concern is to define and build models that represent different scenarios of multimedia streaming systems. We focus on simulating different complicated network topologies as a trial to approach the reality. We measure end-to-end delay, jitter and packet loss as an indication of accepted synchronized multimedia sources of multiparty communication session. In this paper we elaborate on bandwidth utilization as a main measured parameter in this area of multimedia streaming.

Our simulation study is performed using the Network Simulator ns-2, [6], on a Linux platform. Network performance is examined from the unicast as well as the multicast routing point of view, with emphasis on network behavior in case of bottleneck links and link failure transitions.

The sample network topologies used are also examined with and without the support of Quality of Service architectures that supports Differentiated Services (DiffServ) and Multi-Protocol Label Switching (MPLS). However, in this paper we elaborate on our studies with multicast routing and we reserve the performance evaluation of QoS architecture to future studies.

One of our sample network topologies that we used in this study, Figure 2, consists of core and edge routers that connect four source nodes as traffic generators and three destination nodes.

One multicast group is defined with the source node S1 as a parent node and two children nodes at the destination nodes D1 and D2. The other three unicast sessions S2-D2, S3-D2, and S4-D3 are maintained. The Dense Mode (DM) multicast routing protocol is configured with Distance Vector Multicast Routing Protocol (DVMRP), [5].

We have run different simulation scenarios to examine the behavior of different multimedia traffic sources.

The sample network topology was examined with all traffic sources generating Constant Bit Rate (CBR) traffic with the same transmission rate.

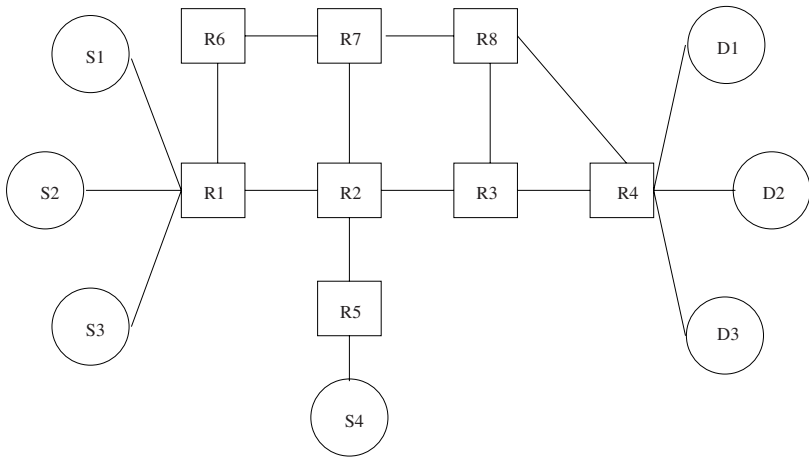


Fig. 2. Sample Network Topology

The experiments are repeated with CBR traffic of different transmission rates. The network topology is also examined with Real Audio multicast traffic sources connected to S1 and the other three unicast traffic sources are generating CBR traffic with varying transmission rate.

We have used the MPEG4 and H.263 Video trace files generated at the Technical University Berlin, [4].

Due to technical limitations of the simulator, in the scenarios demonstrated in this paper we simulated either the video or audio traffic sources using adjusted CBR traffic.

4 Analysis and Discussion

The sample network topology, Figure 2, was tested with Real Audio multicast traffic source connected to S1 and the other three unicast traffic sources generate CBR traffic with transmission rate of 5Mbps. This network setup has shown that the multicast traffic source as well as the fourth traffic source are fully discarded in the steady state. The second and third traffic sources were transmitted, with each source of 5Mbps, sharing the limited bandwidth of the bottleneck link, Figure 3.

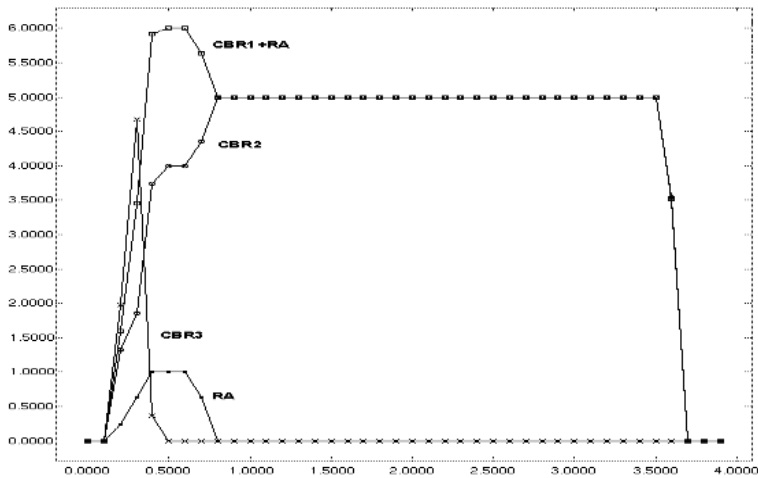


Fig. 3. Bandwidth Utilization with Time of Real Audio and CBR Traffic

This experiment shows that unicast routing is dominating. The multicast routing consumes longer time to establish the parent-child relationship. The unicast connection is faster to be established and, therefore, in case of limited bandwidth, the multicast connection suffers packets drop.

The experiment was repeated with reducing the traffic sources S2 and S3 to generate CBR traffic of only 2Mbps, while S1 keeps generating Real Audio traffic of 1Mbps and S4 keeps generating CBR traffic of 5Mbps. In this case the total bandwidth injected through the sample network topology is 10Mbps. In

this experiment, all flows were delivered and nothing was discarded regardless of the bottleneck link. The multicast flow manages to reach its destinations D1 and D2 successfully. Only the fourth flow suffers a bandwidth reduction to 4Mbps instead of the requested 5Mbps.

It is worth to mention that changing the location of the bottleneck link changes the behavior of the traffic sources all over the route between sources and destinations. The experiment was repeated to test the network behavior in case of link down/up transition. The whole traffic traversing this link was fully discarded until the next up state of the link. Neither the multicast nor the unicast flows are capable of using another route. Therefore, multicast routing has to be established in addition to the MPLS architecture. We reserve this test for future studies.

The experiment was repeated with all link capacities are similar, i.e., configured of 10Mbps bandwidth. The transmission rate of S4 is raised to 10Mbps. The multicast Real Audio traffic is configured to start to transmit earlier than the other three unicast flows. This test shows full bandwidth utilization of the first three flows. The fourth flow was the one suffering reduction to use the rest of the available bandwidth, Figure 4.

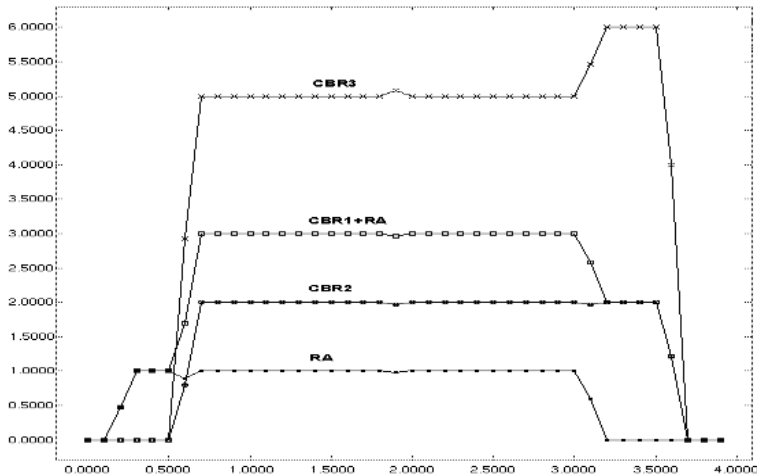


Fig. 4. Bandwidth Degradation with Time

In case of all traffic flows start together, the fourth flow manages to have a pulse of 10Mbps flow rate, only as a transient phase, while all other flows are suddenly dropped, then things stabilize to the steady transmission of the case above. This means flow 3 falls down to 5Mbps, while others are transmitted with the full requested rate.

Exchanging the transmission rate of S2 and S4, i.e., S2 10Mbps, and S4 2Mbps, this affects all other flows dramatically. The simulation study shows that all traffic flows are served according to first in first out. This means the bandwidth allocated to flows cannot be downgraded. The next flows are served with the available bandwidth. The second flow consumes all the capacity of the sample network topology. It is a unicast flow so it dominates the multicast session and the next two unicast sessions. This behavior does not fully comply with the reality of the best-effort IP network where all the sources are expected to share the available bandwidth with suffering of delay and packet loss. However, with QoS-enabled network, it is supposed to be capable of allocating bandwidth to the marked flow according to the Service Level Agreement (SLA) between the customer and the service provider. We reserve this test for future studies.

In summary, multicast routing helps to improve the network performance by reducing the bandwidth consumption of one to many connection. However, it is not capable by itself to solve the problems of link utilization and rerouting techniques. It has to be considered with the MPLS and DiffServ architectures. All of these techniques are proposed to solve the problem of limited link capacities of the existing infrastructure. Improving the link capacities solve the performance problems from the routing point of view.

5 Conclusion

In this paper we present the current version of our developed system of event-scheduling multimedia presentation. The main objective to develop the system is to enable fulfilling the user needs to formulate interactively his own presentation according to his pace and wish. Our user is mainly the university student or researcher. We also introduce our associated simulation study of a multiparty synchronized streaming system.

We have seen that the multicast routing can improve the system response time but in some network scenarios the unicast routing dominates. This means in these scenarios the multicast connections suffer serious packet drops that affect dramatically the quality of the received multimedia traffic.

According to our experience, we believe that we are the first trial to simulate the user-driven multimedia sessions. This task is challenging and needs a great effort to define the framework and architecture. The simulation study was done using the available tools and algorithms. There is a need to implement new codecs for network simulator. Moreover, there is a great need to combine the efforts to model and simulate various topologies and network scenarios that are not only focusing on the routing and networking layers, but it has to move to simulate the complexities of the user interaction with the higher layers.

In summary, our system is highly promising from the performance point of view as well as the easiness to implement and the reproducibility in different forms as well as different time-scheduling.

Acknowledgement. We would like to thank our colleagues at the department, Ms. P. Gonzales, Mr. I. Akkaya, Mr. B. Eren, Ms. Y. Günter, Mr. A. Schmeiser, and Mr. M. Rabel.

References

1. N. Abdel-Baki, B. Aumann, H. P. Großmann: Multimedia Synchronization Based on CDP and DEFSM. Proceedings IEEE MTAC2001, Multimedia Technology and Applications Conference, Irvine, California (2001)
2. N. Abdel-Baki, B. Aumann, H. P. Großmann: Analyzing Multimedia Streaming in a Distributed Environment. Proceedings IEEE ECUMN2002, 2nd European Conference on Universal Multiservice Networks, Colmar, France (2002)
3. N. Abdel-Baki, E. Pérez-Soler, B. Aumann, H. P. Großmann: A Simplified Design and Implementation of a Multimedia Streaming System. Proceedings IEEE ICT'2003, International Conference on Telecommunications, Papeete, Tahiti, French Polynesia (2003)
4. F. H. P. Fitzek, M. Reisslein: MPEG-4 and H.263 Video Traces for Network Performance Evaluation. Technical Report, TKN-00-06, Technical University Berlin, <http://www-tnk.ee.tu-berlin.de/fitzek/TRACE/pub.html> (2000)
5. J. Moy: Multicast Extensions to OSPF. RFC1075 (1991)
6. The Network Simulator ns-2, <http://www.isi.edu/nsnam/ns/>

Applying Linux High-Availability and Load Balancing Servers for Video-on-Demand (VOD) Systems*

Chao-Tung Yang¹, Ko-Tzu Wang¹, Kuan-Ching Li², and Liang-Teh Lee³

¹ High Performance Computing Laboratory, Dept. of Computer Science and Information Engineering, Tunghai University, Taichung 407, Taiwan ROC
ctyang@mail.thu.edu.tw

² Parallel and Distributed Processing Center, Dept. of Computer Science and Information Management, Providence University, Taichung 433, Taiwan ROC
kuancli@pu.edu.tw

³ Dept. of Computer Science and Engineering, Tatung University, Taipei 104, Taiwan ROC
ltlee@cse.ttu.edu.tw

Abstract. In this research paper, we integrate and implement High-Availability (HA) and Load-Balancing technologies to clusters of workstations, increasing both the availability and scalability of services and resources in these systems. The cluster of workstations is commonly built and used as web-based VOD servers. Thus, this paper presents the hardware and software configurations of cluster systems working as VOD servers.

Keywords: VOD servers, Cluster Computing Systems, High-Availability, Load Balancing.

1 Introduction

Supercomputers are systems that lead the world in processing capacity and power. Many of design tricks that enabled past supercomputers to outperform all other systems are now incorporated into personal computers, making a single modern desktop PC of today more powerful than a 15-year old supercomputer. Because of problems carried out by supercomputers that can easily be split up into smaller parts to be worked on simultaneously, traditional supercomputers can often be replaced by parallel-processing systems “*clusters of workstations*”, several individual personal machines programmed and interconnected among themselves to act as one computer.

A PC-based cluster is basically a computer system that interconnects two or more individual computers or systems (often called “computing nodes”) that cooperate with each other in order to execute applications. Common end-goals for building and running applications on a cluster system includes: an increase in reliability, load distribution and achievement of high-performance. There are two basic types of cluster systems: high-availability cluster systems and load balancing cluster systems.

Nowadays, Video-on-Demand (VOD) includes a diverse set of services and opportunities. Today’s technology allows telecommunication network operators to offer services

* This research was supported in part by National Science Council, Taiwan, under grant no. NSC92-2213-E-126-006.

such as home shopping, games, and movies on demand. These services should have a competitive price comparing to the video rental, and customers do not need to travel for these services. These possibilities have been reached with the development of the telecommunication and electronic industry. The capacity of a hard disk has doubled almost every year at a nearly-constant cost. The useful compression ratio for video has been increased considerably. MPEG format video can be transported at few megabits per second rate. Video-on-Demand (VOD) is a subscriber video service where customers can interactively select his/her choice from a collection of alternatives. As the underlying technologies are relatively new, VOD still lacks a universal standardization. Nevertheless, many research institutes and commercial organizations have established de-facto standards and consequently, there are many operational VOD-related services available today.

In this paper, the technologies of High-Availability and Load Balancing are combined and implemented into a PC-based cluster system, with the goal of increasing both the availability and scalability of services and resources when using this type of computer systems. By combining high-availability and fail-over clustering solutions, it is provided robust performance and virtually non-existent downtime. Combo clusters are the perfect solution for ISPs and network applications in which continuous uptime is critical. Still in this research paper, we show that the VOD system using cluster of workstations proposed in our research achieves high response time and non-disruption of service.

2 Background

Increasing critical commercial applications are developed over the Internet, providing highly available services that are increasingly important. One of the most important advantages of a clustered system is that its hardware and software are implemented with redundancy. High availability can be provided by detecting computing node or daemon failures and reconfiguring the system appropriately, so that the workload can be taken over by the remaining computing nodes of the cluster system.

In fact, high availability is a quite big field. An elegant highly available system may have a reliable group communication sub-system, membership management, quorum sub-systems, concurrent control sub-system and other modules. There are several researches to be done. Though, in this meantime, we can use some existing software to construct highly available Linux Virtual Server (LVS) systems.

Load balancing clusters integrate multiple systems that share the load of incoming requests in an equitably distributed manner [4,5]. The systems do not work together on a single process, but rather handle incoming requests independent one of another. This type of cluster is especially suited to ISPs and e-commerce applications that require real-time resolution of several simultaneous incoming requests. Additionally, to organize a cluster to be scalable, it must ensure that each server is fully utilized. The standard technique for accomplishing this is Load Balancing. The basic idea behind load balancing is, by distributing the load proportionally among all the servers in the cluster system, the servers can each run at full capacity, while all requests receive the lowest possible response time. This is the type of cluster system that distributes incoming traffic or resource requests among multiple computing nodes running the same programs. Every node in the cluster

is able to handle requests. If a computing node fails, requests are redistributed among remaining available computing nodes. This type of request distribution is very useful in a web hosting environment. The system architecture of load balancing servers is shown in Figure 1-Left.

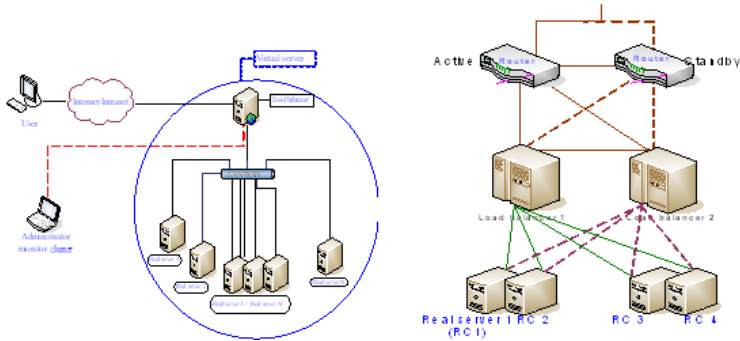


Fig. 1. Left. System architecture of Load-Balancing servers. **Right.** System architecture of High-Availability servers.

Fail-over clustering technique [1,4,5] allows Network Administrators to improve significantly quality of service levels for practically every TCP/IP based network service, such as WWW, Mail, News, and FTP. Unlike distributed processing clusters, high-availability clusters seamlessly and transparently integrate existing stand-alone, non-cluster aware applications, together into a single virtual network, providing the architectural framework necessary to allow the network to continuously and effortlessly grow, to meet performance and reliability demands. This type of servers is shown in Figure 1-Right. This technology aims at achieving uninterrupted availability of services and resources through the use of redundancy built into the system. The general idea is that, if one computer node in the cluster fails, applications or services that were running on that node, it automatically “fail-overs” to an available running node. The types of applications that typically use this type of clusters can be listed as mission-critical database, mail, and file and application servers.

The file system used in our VOD system is parallel virtual file system (PVFS). The PVFS project has been conducted jointly between The Parallel Architecture Research Laboratory at Clemson University and The Mathematics and Computer Science Division at Argonne National Laboratory [6,7,8,9,10,11]. The PVFS is an effort to provide a parallel file system for PC-based clusters. As a parallel file system, PVFS provides a global name space, striping data across multiple I/O nodes, and multiple user interfaces as shown in Figure 2-Left. The system is implemented at user level, so no kernel modifications are necessary to install or run such system. All communications are performed using TCP/IP, so no additional message passing libraries are needed, and support is included for using existing binaries on PVFS files.

3 The Proposed System

Video-On-Demand (VOD) systems are designed by integrating high-availability and load-balancing servers on PC-based clusters has been discussed in [2,3,12,13,14]. In this section, it is described the design of the proposed system, its implementation, and the rationale behind it.

Figures 2-Right, 3-Left and 3-Right show the system architecture and hierarchy of a typical VOD system. The main objective of VOD system is to simplify the storage of thousands of hours of multimedia material, integrating high-capacity, high-latency, tertiary storage systems, called archive servers, with low-capacity, rapid-access, secondary storage systems, called video file servers. We have developed a storage policy (i.e., cache policy) to determine which files need to be migrated to the video file server, in order to reduce the expected access time when a user requests material. VOD system establishes a framework where it is possible to implement different cache policies that permits faster answers to the clients based on the media access patterns.

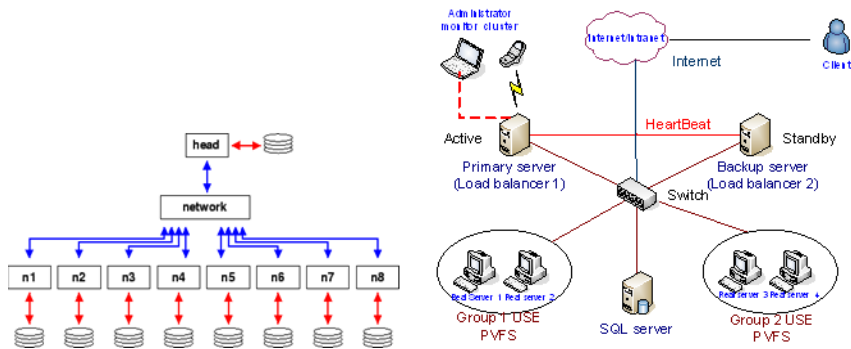


Fig. 2. Left. PVFS consists of multiple HDs on PCs. Right. System architecture of a typical VOD system.

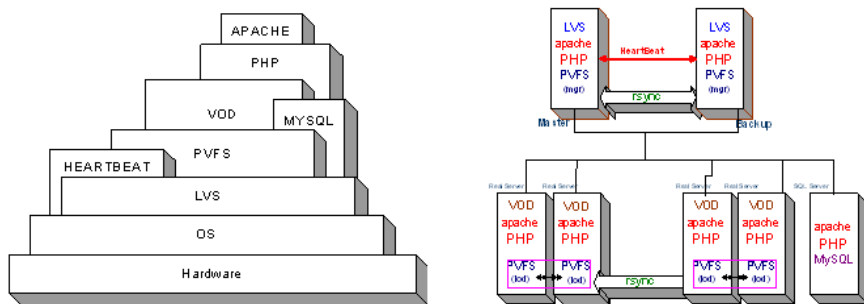


Fig. 3. Left. Hierarchy of a VOD architecture. Right. VOD system.

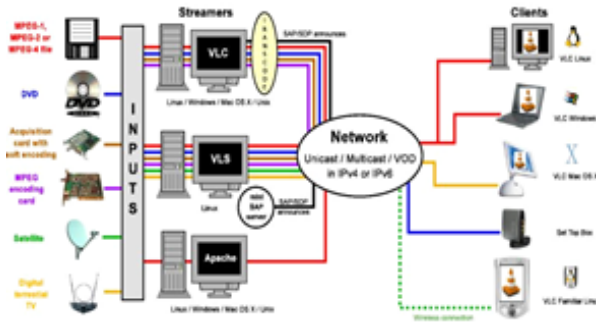


Fig. 4. System architecture of VideoLAN.

If we have several clients trying to access stored media, several video file servers that can play media, one or more archive servers that integrate tertiary storage into the architecture by using PVFS and a database that stores meta-data using MySQL needed to identify stored material and locate desired material. See figure 4 for more details.

The second goal of VOD system is to support media assets with different video, audio, image, and text formats, and thus, avoiding the user to deal with multiple formats inherent in multimedia material. Users can add new data types and formats by providing directives that describe how to treat new type or format. For example, a user can have different versions of the same video shot (e.g., AVI, MPEG, and RealVideo formats), but he wants to handle these files as only one object.

The VideoLAN [12] is selected and used in our VOD system as shown in Figure 4. This VideoLAN solution includes:

- VLS (VideoLAN Server), which can stream MPEG-1, MPEG-2 and MPEG-4 files, DVDs, digital satellite channels, digital terrestrial television channels and live videos on the network in unicast or multicast,
- VLC (initially VideoLAN Client), which can be used as a server to stream MPEG-1, MPEG-2 and MPEG-4 files, DVDs and live videos on the network in unicast or multicast; or used as a client to receive, decode and display MPEG streams under multiple operating systems.

The following libraries are installed in each client:

- Libdvdcss: to be able to read encrypted DVDs,
- Libdvdpplay: to have DVD menu navigation,
- Libdvbpsi: to be able to read from the network,
- a52dec: to be able to decode the AC3 (i.e. A52) sound format often used in DVDs,
- ffmpeg, libmad and faad2: to read MPEG-4 / DivX files,
- libogg and libvorbis: to read Ogg Vorbis files.

On the client side, the user can input commands to initiate the screen in order to view the video in Linux OS environments. See Figures 5-Left and 5-Right for screen visualization.

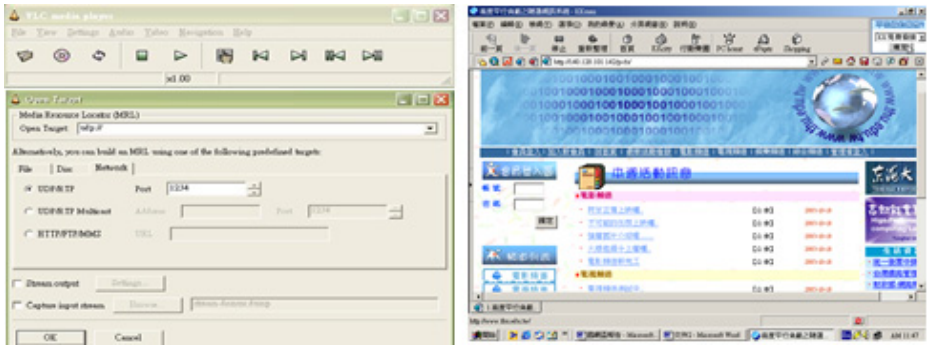


Fig. 5. Left. VLC screen view Right. VOD system portal.

As a VOD server, you need a running Web server. For example, you can use a Linux server running Apache. Make your MPEG-1, MPEG-2 or MPEG-4/DivX files available to the clients on the Web server. For example, we have a Web server that DNS name is local server. In this server, we put a MPEG file video1.xyz that is available to clients at the URL `http://localserver/video1.xyz` and use the following command `[root@fs1/root]# vlc -vvv video1.mpeg --sout udp:User's IP`

The third objective of the VOD system is to permit easy control (e.g., browsing, addition, modification, and deletion) of material using a well-known desktop interface metaphor as shown in Figures 6-Left and 6-Right. In addition, the system should allow remote users to access the system using the Internet. The goals have been achieved by integrating access to the system into a web-based Graphical User Interface (GUI).

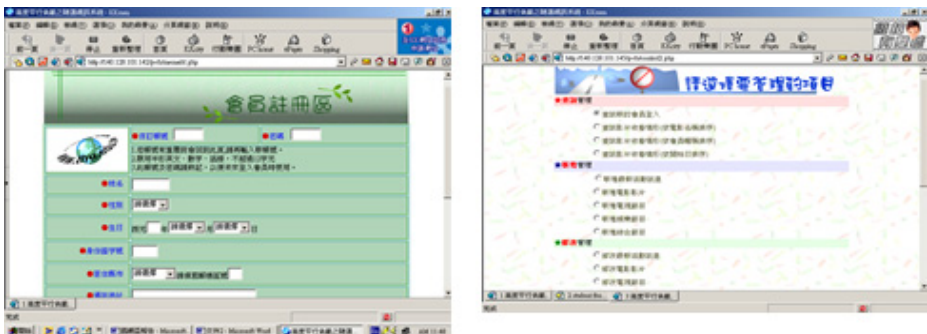


Fig. 6. Left. VOD system register function. Right. VOD system monitor and control.

Finally, the system provides access control through a remote identification and authentication service. It implements a policy to relate ownership and access control and therefore, enforce access restrictions. In this way, it is possible to create management services such as copyright protection and billing.

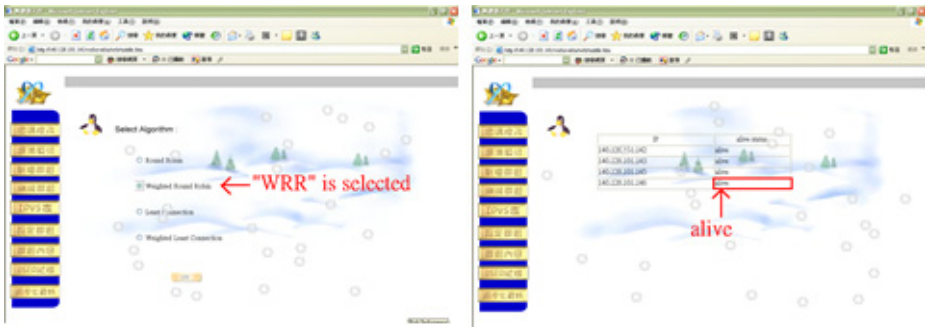


Fig. 7. Left. Scheduling selection function. **Right.** VOD system status menu.

Our VOD system offers three types of load-balancing cluster solutions as shown in Figures 7-Left and 7-Right. Each utilizes the highest-quality, performance-optimized system of load balancing hardware, software and components.

- *Fewest Connections*: This technique routes requests to the server that is currently handling the smallest number of requests/connections. For example, if Server 1 is currently handling 20 requests, and Server 2 is currently handling 10 requests/connections, the next request/connection will be automatically routed to Server 2, since that server currently has the least number of active connections/requests.
- *Round Robin*: This technique routes requests to the next available server on a rotating basis. For example, incoming requests/connections are routed to Server 1, then Server 2, and finally Server 3 before starting again with Server 1.
- *Weighted Fair*: This technique routes requests to servers based upon their current request load and their performance capabilities. For example, If Server 1 is four times faster at handling requests than Server 2; the administrator factors this difference by routing 4 times as many performance loads to Server 1 than Server 2.

The experimental environment is built using the following hardware and software configurations on master and backup nodes: CPU Intel PIII 1GHz * 2 CPUs, 512MB DDR memory, HD 30GB, OS RedHat Linux 8.0 and 9.0, application software LVS (Linux Virtual Server), The IPVS Netfilter module for kernel 2.4, HA (High Availability)-HeartBeat, WEB Apache and PHP, also RSYNC, IPVSADM, SUPER and REALCHK.

The hardware and software configurations to work as redundancy are listed as: CPU Intel Celeron 900 MHz, 256MB memory, HD 30GB, OS RedHat Linux 8.0 and 9.0, application software LVS (Linux Virtual Server), the IPVS Netfilter module for kernel 2.4, PVFS v1.6.0, pvfs-kernel v1.6.0, VOD vlc-0.6.1-1.i386.rpm and FFmpeg version 0.4.8, WEB Apache, PHP, and also, SUPER and RSYNC.

The hardware and software configuration of database node (SQL Server) is: CPU Intel Celeron 900 MHz, 256MB DDR memory, HD 30GB, OS RedHat Linux 9.0, and the application software used is MySQL database.

4 Conclusion

High-Availability (HA) and load balancing clusters are used when it is critical the content or service to be available and/or processed as fast as possible. Internet Service Providers or E-commerce web sites services require high availability, load balancing and scalability.

In this paper, we integrated the technologies of High-Availability and Load Balancing in clusters of workstations, by combining features of both of the cluster types, increasing both the availability and scalability of services and resources. Nowadays, PC-based clusters with this technology are commonly used for web-based VOD servers.

References

1. LVS (Linux Virtual Server), <http://www.linuxvirtualserver.org/>
2. Apache Server, <http://www.apache.org/>
3. PHP programming, <http://www.php.net/>
4. HA (High Availability), <http://www.linuxvirtualserver.org/HighAvailability.html>
5. Linux HA, <http://www.linux-ha.org/>
6. The OpenGFS Project, <http://opengfs.org/>
7. Sistina's Global File System (GFS), http://www.sistina.com/products_gfs.htm
8. GPFS, <http://www-124.ibm.com/developerworks/opensource/gpfs/>
9. General Parallel File System for Linux,
<http://www-1.ibm.com/servers/eserver/clusters/software/gpfs.html>
10. The Parallel Virtual File System Project, <http://parlweb.parl.clemson.edu/pvfs/>
11. Journal File System Technology for Linux, <http://www-124.ibm.com/developerworks/oss/jfs/>
12. VideoLan - Open Source Video Streaming Solution, <http://www.videolan.org/>
13. Sourceforge, <http://ffmpeg.sourceforge.net/index.org.html>
14. MySQL database, <http://www.mysql.com/>

Indexing Issues in Supporting Similarity Searching*

Hanan Samet

Computer Science Department, Center for Automation Research, Institute for Advanced
Computer Studies, University of Maryland, College Park, Maryland 20742
hjs@cs.umd.edu, www.cs.umd.edu/~hjs

Abstract. Indexing issues that arise in the support of similarity searching are presented. This includes a discussion of the curse of dimensionality, as well as multidimensional indexing, distance-based indexing, dimension reduction, and embedding methods.

1 Introduction

The representation of multidimensional points and objects, and the development of appropriate indexing methods that enable them to be retrieved efficiently is a well-studied subject (e.g., [1,2]). Most of these methods were designed for use in application domains where the data usually has a spatial component which has a relatively low dimension. Examples of such application domains include geographic information systems (GIS), spatial databases, solid modeling, computer vision, computational geometry, and robotics. However, there are many application domains where the data is of considerably higher dimensionality, and is not necessarily spatial. This is especially true in multimedia databases where the data is a set of objects and the high dimensionality is a direct result of trying to describe the objects via a collection of features (also known as a *feature vector*). In the case of images, examples of features include color, color moments, textures, shape descriptions, etc. expressed using scalar values. The goal in these applications is often expressed more generally as one of the following:

1. Find objects whose feature values fall within a given range or where the distance from some query object falls into a certain range (range queries).
2. Find objects whose features have values similar to those of a given query object or set of query objects (nearest neighbor queries).

These queries are collectively referred to as *similarity searching*, and the issues involved in supporting them is the subject of this paper, which is organized as follows. Section 2 mentions the use of Voronoi diagrams, while Section 3 describes the curse of dimensionality. Sections 4 and 5 discuss multidimensional indexing and distance-based indexing, respectively, while Section 6 briefly touches on dimension reduction and embedding methods. Concluding remarks are drawn in Section 7.

* This work was supported in part by the National Science Foundation under grants EIA-99-00268, IIS-00-86162, and EIA-00-91474, and Microsoft Research.

2 Voronoi Diagrams

An apparently straightforward solution to finding the nearest neighbor is to compute a Voronoi diagram (e.g., [3]) for the data points (i.e., a partition of the space into regions where all points in the region are closer to the region's associated data point than to any other data point), and then locate the Voronoi region corresponding to the query point. The problem with this solution is that the combinatorial complexity of the Voronoi diagram in high dimensions is prohibitive — that is, it grows exponentially with its dimension k so that for N points, the time to build and the space requirements can grow as rapidly as $\Theta(N^{k/2})$ [3]. This renders its applicability moot.

3 Curse of Dimensionality

The above is typical of the problems that we must face when dealing with high-dimensional data. Generally speaking, multidimensional queries become increasingly more difficult as the dimensionality increases. The problem is characterized as the *curse of dimensionality*. This term was coined by Bellman [4] to indicate that the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with the number of variables (i.e., dimensions) that comprise it. For similarity searching (i.e., finding nearest neighbors), this means that the number of objects (i.e., points) in the data set that need to be examined in deriving the estimate grows exponentially with the underlying dimension.

The curse of dimensionality has a direct bearing on similarity searching in high dimensions as it raises the issue of whether or not nearest neighbor searching is even meaningful in such a domain. In particular, letting d denote a distance function which need not necessarily be a metric, Beyer et al. [5] point out that nearest neighbor searching is not meaningful when the ratio of the variance of the distance between two random points p and q , drawn from the data and query distributions, and the expected distance between them converges to zero as the dimension k goes to infinity — that is,

$$\lim_{k \rightarrow \infty} \frac{\text{Variance}[d(p, q)]}{\text{Expected}[d(p, q)]} = 0.$$

In other words, the distance to the nearest neighbor and the distance to the farthest neighbor tend to converge as the dimension increases. Formally, Beyer et al. demonstrate that when the data and query distributions satisfy this ratio, the probability that the farthest neighbor distance is smaller than $1 + \epsilon$ of the nearest neighbor distance is 1 in the limit as the dimension k goes to infinity and ϵ is a positive value. For example, they show that this ratio holds whenever the coordinate values of the data and the query point are independent and identically distributed as is the case when they are both drawn from a uniform distribution.

Assuming that d is a distance metric and hence that the triangle inequality holds, an alternative way of looking at the curse of dimensionality is to observe that when dealing with high-dimensional data, the probability density function (analogous to a histogram) of the distances of the various elements is more concentrated and has a larger mean value. This means that similarity searching algorithms will have to perform more work.

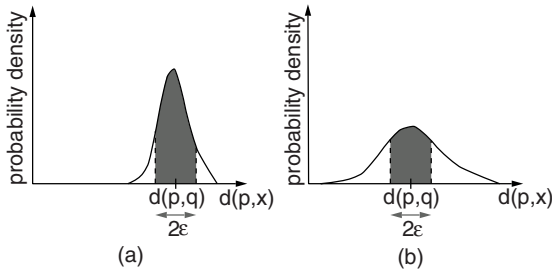


Fig. 1. A probability density function (analogous to a histogram) of the distances $d(p,x)$ with the shaded area corresponding to $|d(q,p) - d(p,x)| \leq \epsilon$. (a) indicates a density function where the distance values have a small variation, while (b) indicates a more uniform distribution of distance values thereby resulting in a more effective use of the triangle inequality to prune objects from consideration as satisfying the range search query.

In the worst case, we have the situation where $d(x,x) = 0$ and $d(x,y) = 1$ for all $y \neq x$, which means that a similarity query must compare the query object with every object of the set. One way to see why more concentrated probability densities lead to more complex similarity searching is to observe that this means that the triangle inequality cannot be used so often to eliminate objects from consideration. In particular, the triangle inequality implies that every element x such that $|d(q,p) - d(p,x)| > \epsilon$ cannot be at a distance of ϵ or less from q (i.e., from $d(q,p) \leq d(p,x) + d(x,q)$). Thus if we examine the probability density function of $d(p,x)$ (i.e., on the horizontal axis), we find that when ϵ is small while the probability density function is large at $d(p,q)$, then the probability of eliminating an element from consideration via the use of the triangle inequality is the remaining area under the curve, which is quite small (see Figure 1a in contrast to Figure 1b where the density function of the distances is more uniform).

These observations mean that nearest neighbor searching may be quite inefficient as it is difficult to differentiate between the nearest neighbor and the other elements. Moreover, seemingly appropriate indexing methods, such as k-d trees [6] and R-trees [7] which are designed to make it easier to avoid examining irrelevant elements, may not be of use in this case. In fact, the experiments of Beyer et al. [5] show that the curse of dimensionality becomes noticeable for dimensions as low as 10 to 15 for the uniform distribution. The only saving grace is that real world high-dimensional data (say of dimension k) is not likely to be uniformly distributed as their volume is much smaller than $O(c^k)$ for some small constant $c > 2$. Thus we can go on with our discussion despite the apparent pall of the curse of dimensionality which tends to cast a shadow on any arguments or analyses that are based on uniformly-distributed data or queries.

4 Multidimensional Indexing

Assuming that the curse of dimensionality does not come into play, query responses are facilitated by sorting the objects on the basis of some of their feature values and building appropriate indexes. The high-dimensional feature space is indexed using some

multidimensional data structure (termed *multidimensional indexing*) with appropriate modifications to fit the high-dimensional problem environment. Similarity search which finds objects similar to a target object can be performed with a range search or a nearest neighbor search in the multidimensional data structure. However, unlike applications in spatial databases where the distance between two objects is usually Euclidean, this is not necessarily the case in the high-dimensional feature space where the distance function may even vary from query to query on the same feature (e.g., [8]).

Searching in high-dimensional spaces is time-consuming. Performing range queries in high dimensions is considerably easier, from the standpoint of computational complexity, than performing similarity queries as range queries do not involve the computation of distance. In particular, searches through an indexed space usually involve relatively simple comparison tests. However, if we have to examine all of the index nodes, then the process is again time-consuming. In contrast, computing similarity in terms of nearest neighbor search makes use of distance and the process of computing the distance can be computationally complex. For example, computing the Euclidean distance between two points in a high-dimensional space, say d , requires d multiplication operations and $d - 1$ addition operations, as well as a square root operation (which can be omitted). Note also that computing similarity requires the definition of what it means for two objects to be similar, which is not always so obvious.

5 Distance-Based Indexing

Often, the only information that we have available is a distance function that indicates the degree of similarity (or dis-similarity) between all pairs of the N given objects. Usually the distance function d is required to obey the triangle inequality, be non-negative, and be symmetric, in which case it is known as a *metric* and also referred to as a *distance metric*. Sometimes, the degree of similarity is expressed by use of a similarity matrix which contains interobject distance values, for all possible pairs of the N objects

Given a distance function, we usually index the data (i.e., objects) with respect to their distance from a few selected objects. We use the term *distance-based indexing* to describe such methods (e.g., [9]). A number of such methods have been proposed over the past few decades, some of the earliest being due to Burkhard and Keller [10]. These methods generally assume that we are given a finite set S of N objects and a distance metric d indicating the distance values between them (collectively termed a *finite metric space*) Typical of distance-based indexing structures are *metric trees* [11,12], which are binary trees that result in recursively partitioning a data set into two subsets at each node. Uhlmann [12] identified two basic partitioning schemes, *ball partitioning* and *generalized hyperplane partitioning*.

In ball partitioning, the data set is partitioned based on distances from one distinguished object, sometimes called a *vantage point* [13], into the subset that is inside and the subset that is outside a ball around the object (e.g., see Figure 2a). In generalized hyperplane partitioning, two distinguished objects p_1 and p_2 are chosen and the data set is partitioned based on which of the two distinguished objects is the closest — that is, all the objects in subset A are closer to p_1 than to p_2 , while the objects in subset B are closer to p_2 (e.g., see Figure 2b). The asymmetry of ball partitioning (which is

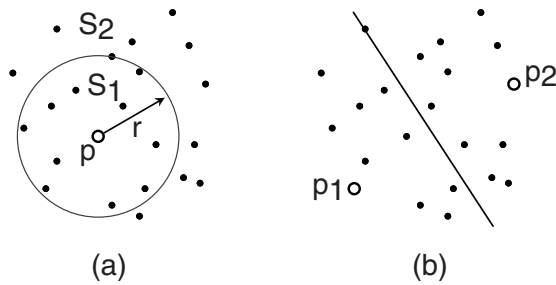


Fig. 2. Possible top-level partitionings of a set of objects (depicted as two-dimensional points) in a metric tree using (a) ball partitioning and (b) generalized hyperplane partitioning.

evident from Figure 2a) is a potential drawback of this method as the outer shell tends to be very narrow for metric spaces typically used in similarity search (e.g., see [14]). In contrast, generalized hyperplane partitioning is more symmetric, in that both partitions form a “ball” around an object (see Figure 2b). The vp-tree [13] is an example of a ball partitioning tree while the gh-tree [12] is an example of a generalized hyperplane partitioning tree.

An alternative way of distinguishing between some of the different distance-based indexing methods is on the basis of whether they are pivot-based or clustering-based (e.g., [15]). Pivot-based methods choose a subset of the objects in the data set to serve as distinguished objects, termed *pivot objects* (or more generally *pivots*), and classify the remaining objects in terms of their distances from the pivot objects. Pivot-based similarity searching algorithms make use of the known distances from the objects to different pivot objects to reduce the number of distance computations involving the query object that will be needed to respond to the query. The pivot objects, assuming without loss of generality that there are k of them, can often be viewed as coordinates in a k -dimensional space and the result of the distance computation for object x is equivalent to a mapping of x to a point $(x_0, x_1, \dots, x_{k-1})$ where coordinate value x_i is the distance $d(x, p_i)$ of x from pivot p_i . The result is similar to embedding methods which are discussed further below. Ball partitioning methods are examples of pivot-based methods. In addition, methods that make use of distance matrices which contain precomputed distances between some or all of the objects in the data set (e.g., [16]) are also examples of pivot-based methods. Note that distance matrix methods differ from ball partitioning methods in that they do not form a hierarchical partitioning of the data set.

Clustering-based methods partition the underlying data set into spatial-like zones called *clusters* that are based on proximity to a distinguished object known as the *cluster center*. In particular, once a set of cluster centers has been chosen, the objects that are associated with each cluster center c are those that are closer to c than to any other cluster center. Although the cluster centers play a similar role as the pivot objects, the principal difference is that an object o is associated with a particular pivot p on the basis of the distance from o to p and not because p is the closest pivot to o , which would be the case if p was a cluster center. Generalized-hyperplane partitioning methods are examples

of clustering-based methods. The *sa-tree* [17,18], inspired by the Voronoi diagram, is another example of a clustering-based method. It records a portion of the Delaunay graph of the data set, which is a graph whose vertices are the Voronoi cells, with edges between adjacent cells. Although many of the clustering-based methods are hierarchical, this need not necessarily be the case.

It is interesting to observe that both pivot-based and clustering-based methods achieve a partitioning of the underlying data set into spatial-like zones. However, the difference is that the boundaries of the zones are more well-defined in the case of pivot-based methods as they can be expressed explicitly using a small number of objects and a known distance value. In contrast, in the case of clustering-based methods, the boundaries of the zones are usually expressed implicitly in terms of the cluster centers, instead of explicitly, which may require quite a bit of computation to determine. In fact, very often, the boundaries cannot be expressed explicitly as, for example, in the case of an arbitrary metric space (in contrast to a Euclidean space) where we do not have a direct representation of the ‘generalized hyperplane’ that separates the two partitions.

The advantage of distance-based indexing methods is that distance computations are used to build the index, but once the index has been built, similarity queries can often be performed with a significantly lower number of distance computations than a sequential scan of the entire dataset. Of course, in situations where we may want to apply several different distance metrics, then the drawback of the distance-based indexing techniques is that they require that the index be rebuilt for each different distance metric, which may be nontrivial. This is not the case for the multidimensional indexing methods which have the advantage of supporting arbitrary distance metrics (however, this comparison is not entirely fair, since the assumption, when using distance-based indexing, is that often we do not have any feature values as for example in DNA sequences).

6 Dimension Reduction and Embedding Methods

There are many problems with indexing high-dimensional data. In particular, it can be shown that, assuming uniformly-distributed high-dimensional data, most of the data lies at or near the boundary of the data space [19] (e.g., for 20 dimensions, 98.85% of data lies in the outermost 10% of the hypercube of the data space). Therefore, only rarely is the data volume so high that every dimension is split even once when using an index such as a k-d tree. Thus a typical query region often overlaps all of the leaf node regions of the index which means that the cost of performing queries using the index is often higher than a sequential scan of the entire data (e.g., [5,20]). In fact, this is another manifestation of the curse of dimensionality.

Nevertheless, the “inherent dimensionality” of a data set is often much lower than the dimensionality of the underlying space. For example, the values of some of the features may be correlated in some way. Thus there has been a considerable amount of interest in techniques to reduce the dimensionality of the data using methods such as Singular Value Decomposition (SVD) [21], Karhunen-Loève Transform (KLT) [22], and Principal Component Analysis (PCA) [22]. Another motivation for the development of many dimension-reduction techniques has been a desire to make use of disk-based spatial indexes which are based on object hierarchies such as members of the R-tree family [7,

23,24]. The performance of these methods decreases with an increase in dimensionality due to the decrease in the fanout of a node of a given capacity since usually the amount of storage needed for the bounding boxes is directly proportional to the dimensionality of the data thereby resulting in longer search paths.

In situations where no features are defined for the objects but only a distance function, there exists an alternative to using distance-based indexes. In particular, methods have been devised for deriving “features” purely based on the inter-object distances [25,26, 27,28]. Thus, given N objects, the goal is to choose a value of k and find a set of N corresponding points in a k -dimensional space so that the distance between the N corresponding points is as close as possible to that given by the distance function for the N objects. In particular, if the methods are contractive (i.e., the distance in the embedding space is always less than the distance in the original space) [29], then we can now index the points using multidimensional data structures while guaranteeing 100% recall (i.e., that we will not miss any objects). These methods are known as *embedding methods* and can also be applied to objects represented by feature vectors as alternatives to the traditional dimension-reduction methods. Not all embedding methods are contractive for all distance metrics.

7 Concluding Remarks

Providing indexing support for similarity searching is an important area where much work remains to be done. Some of the more promising research directions lie in developing techniques to identify the important features in the applications so that the dimension of the problem domain can be reduced thereby enabling us to properly utilize the vast array of existing indexing and nearest neighbor techniques.

References

1. Samet, H.: Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS. Addison-Wesley, Reading, MA (1990)
2. Samet, H.: The Design and Analysis of Spatial Data Structures. Addison-Wesley, Reading, MA (1990)
3. Aurenhammer, F.: Voronoi diagrams — a survey of a fundamental geometric data structure. *ACM Computing Surveys* **23** (1991) 345–405
4. Bellman, R.E.: Adaptive Control Processes. Princeton University Press, Princeton, NJ (1961)
5. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is “nearest neighbor” meaningful? In Beeri, C., Buneman, P., eds.: Proceedings of the 7th International Conference on Database Theory (ICDT’99), Berlin, Germany (1999) 217–235
6. Bentley, J.L.: Multidimensional binary search trees used for associative searching. *Communications of the ACM* **18** (1975) 509–517
7. Guttman, A.: R-trees: a dynamic index structure for spatial searching. In: Proceedings of the ACM SIGMOD Conference, Boston (1984) 47–57
8. Rui, Y., and S. Mehrotra, T.S.H.: Content-based image retrieval with relevance feedback in MARS. In: Proceedings of the 1997 IEEE International Conference on Image Processing, Santa Barbara, CA (1997) 815–818
9. Hjaltonson, G.R., Samet, H.: Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems* **28** (2003) 517–580

10. Burkhard, W.A., Keller, R.: Some approaches to best-match file searching. *Communications of the ACM* **16** (1973) 230–236
11. Uhlmann, J.K.: Metric trees. *Applied Mathematics Letters* **4** (1991) 61–62
12. Uhlmann, J.K.: Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters* **40** (1991) 175–179
13. Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms*, Austin, TX (1993) 311–321
14. Brin, S.: Near neighbor search in large metric spaces. In Dayal, U., Gray, P.M.D., Nishio, S., eds.: *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB)*, Zurich, Switzerland (1995) 574–584
15. Chávez, E., Navarro, G.: An effective clustering algorithm to index high dimensional spaces. In: *Proceedings String Processing and Information Retrieval (SPIRE 2000)*, A Coruña, Spain (2000) 75–86
16. Vidal Ruiz, E.: An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters* **4** (1986) 145–157
17. Navarro, G.: Searching in metric spaces by spatial approximation. *VLDB Journal* **11** (2002) 28–46
18. Hjalton, G.R., Samet, H.: Improved search heuristics for the sa-tree. *Pattern Recognition Letters* **24** (2003) 2785–2795
19. Berchtold, S., Böhm, C., Kriegel, H.P.: Improving the query performance of high-dimensional index structures by bulk-load operations. In: *Advances in Database Technology — EDBT'98, Proceedings of the 6th International Conference on Extending Database Technology*, Valencia, Spain (1998) 216–230
20. Weber, R., Schek, H.J., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In Gupta, A., Shmueli, O., Widom, J., eds.: *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, New York (1998) 194–205
21. Golub, G.H., van Loan, C.F.: *Matrix Computations*. Third edn. Johns Hopkins University Press, Baltimore, MD (1996)
22. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*. Second edn. Academic Press, Boston (1990)
23. Katayama, N., Satoh, S.: The SR-tree: an index structure for high-dimensional nearest neighbor queries. In Peckham, J., ed.: *Proceedings of the ACM SIGMOD Conference*, Tucson, AZ (1997) 369–380
24. White, D.A., Jain, R.: Similarity indexing with the SS-tree. In Su, S.Y.W., ed.: *Proceedings of the 12th IEEE International Conference on Data Engineering*, New Orleans (1996) 516–523
25. Faloutsos, C., Lin, K.I.: FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: *Proceedings of the ACM SIGMOD Conference*, San Jose, CA (1995) 163–174
26. Hristescu, G., Farach-Colton, M.: Cluster-preserving embedding of proteins. Technical report, Rutgers University, Piscataway, NJ (1999)
27. Linial, N., London, E., Rabinovich, Y.: The geometry of graphs and some of its algorithmic applications. *Combinatorica* **15** (1995) 215–245
28. Wang, J.T.L., Wang, X., Lin, K.I., Shasha, D., Shapiro, B.A., Zhang, K.: Evaluating a class of distance-mapping algorithms for data mining and clustering. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA (1999) 307–311
29. Hjalton, G.R., Samet, H.: Properties of embedding methods for similarity searching in metric spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25** (2003) 530–549 Also University of Maryland Computer Science TR-4102.

Efficient Visual Content Retrieval and Mining in Videos

Josef Sivic and Andrew Zisserman

Robotics Research Group, Department of Engineering Science
University of Oxford
<http://www.robots.ox.ac.uk/~vgg>

Abstract. We describe an image representation for objects and scenes consisting of a configuration of viewpoint covariant regions and their descriptors. This representation enables recognition to proceed successfully despite changes in scale, viewpoint, illumination and partial occlusion. Vector quantization of these descriptors then enables efficient matching on the scale of an entire feature film. We show two applications. The first is to efficient object retrieval where the technology of text retrieval, such as inverted file systems, can be employed at run time to return all shots containing the object in a manner, and with a speed, similar to a Google search for text. The object is specified by a user outlining it in an image, and the object is then delineated in the retrieved shots. The second application is to data mining. We obtain the principal objects, characters and scenes in a video by measuring the reoccurrence of these spatial configurations of viewpoint covariant regions. The applications are illustrated on two full length feature films.

1 Introduction and objectives

Identifying an (identical) object in frames of a video is a challenging problem because an object's visual appearance may be very different due to viewpoint, scale and lighting changes, and partial occlusion. However, recently a number of successful approaches [5, 8,9,11,12,17] have been developed in the Computer Vision literature based on a *weak segmentation* of the image. Rather than attempt to 'semantically' segment the image, e.g. into foreground object and background, an image is represented by a set of overlapping (local) regions. The region segmentation, and their descriptors, are built with a controlled degree of invariance to viewpoint and illumination conditions. Recognition of a particular object then proceeds by matching the descriptor vectors which act as 'barcodes' for that object. The result is that objects can be recognized despite significant changes in viewpoint and, due to the multiple local regions, despite partial occlusion since some of the regions will still be visible in such cases. Matches of descriptor vectors can be pre-computed by vector quantizing, and this in turn enables very efficient applications to be built.

In this work we describe two applications: object retrieval in videos, and data mining in videos. In object retrieval the aim is to retrieve those key frames and shots of a video containing a particular object with the ease, speed and accuracy with which Google retrieves text documents (web pages) containing particular words.

The second application is to data mining. The objective is to extract significant objects, characters and scenes in a video by determining their frequency of occurrence.



Fig. 1. Object query example I. (a) Top row: (left) a frame from the movie ‘Groundhog Day’ with a query region in yellow and (right) a close-up of the query region delineating the object of interest. Bottom row: (left) all 1039 detected affine co-variant regions superimposed and (right) close-up of the query region. (b) (left) two retrieved frames with detected region of interest in yellow and (right) a close-up of the images with affine co-variant regions superimposed. These regions match to a subset of the regions shown in (a). Note the significant change in foreshortening and scale between the query image of the object, and the object in the retrieved frames. Querying all the 5,640 keyframes of the entire movie took 0.36 seconds on a 2GHz Pentium.

For example, the principal actors will be mined because their face or clothes will appear often throughout a film. Similarly, a particular set or scene that re-occurs (e.g. Rick’s bar in ‘Casablanca’) will be ranked higher than those that only occur infrequently (e.g. a particular tree by the highway in a road movie).

There are a number of reasons why it is useful to have commonly occurring objects/characters/scenes. First, they provide entry points for visual search in videos and image databases, or for generating a visual thesaurus [3]. Second, they can be used in forming video summaries [1,4,16]. A third application area is in detecting product placements in a film – where frequently occurring logos or labels will be prominent.

The retrieval and data mining methods will be illustrated for the feature length films ‘Groundhog Day’ [Ramis, 1993] and ‘Casablanca’ [Curtiz, 1942].

2 Quantized Viewpoint Invariant Descriptors

We build on work on viewpoint invariant descriptors which has been developed for wide baseline matching [6,8,11,17], object recognition [5,9,10], and image/video retrieval [12,13].

The approach taken in all these cases is to represent an image by a set of overlapping regions, each represented by a vector computed from the region’s appearance. The region segmentation is designed so that the pre-image of the region corresponds to the same surface region, i.e. their shape is not fixed, but automatically adapts based on the underlying image intensities so as to always cover the same physical surface. Note that the regions are computed independently in each image. In short, the segmentation commutes with the viewpoint transformation between images, and such regions are known as *affine covariant* (since the transformation is locally an affinity). Similar descriptors

are computed for all images, and region matches between images are then obtained by, for example, nearest neighbour matching of the descriptor vectors, followed by disambiguating using local spatial coherence or global relationships (such as a homography transformation). This approach has proven very successful for lightly textured scenes, with robustness up to a five fold change in scale reported in [7].

Affine co-variant regions: In this work, two types of affine co-variant regions are computed for each frame. The first is constructed by elliptical shape adaptation about an interest point. The implementation details are given in [8,11]. The second type of region is constructed using the maximally stable procedure of Matas *et al.* [6] where areas are selected from an intensity watershed image segmentation. Both types of regions are represented by ellipses. These are computed at twice the originally detected region size in order for the image appearance to be more discriminating. For a 720×576 pixel video frame the number of regions computed is typically between 1000-2000. An example is shown in figure 1a.

Each elliptical affine covariant region is represented by a 128-dimensional vector using the SIFT descriptor developed by Lowe [5]. Combining the SIFT descriptor with affine covariant regions gives region description vectors which are invariant to affine transformations of the image.

Vector quantized descriptors: The SIFT descriptors are vector quantized using K-means clustering. The clusters are computed from 474 frames of the video, with about 6K clusters for Shape Adapted regions, and about 10K clusters for Maximally Stable regions. All the descriptors for each frame of the video are assigned to the nearest cluster centre to their SIFT descriptor. Vector quantizing brings a huge computational advantage because descriptors in the same clusters are considered matched, and no further matching on individual descriptors is then required. In an analogy with text retrieval these vector quantized descriptors are termed *visual words*: they provide a vocabulary – visual nouns – for representing an object or scene [13].

Stop list: The frequency of occurrence of single words across the whole video (database) is measured, and the top 5% are stopped. This step is inspired by a stop-list in text retrieval applications where poorly discriminating very common words (such as ‘the’) are discarded. In the visual word case the large clusters often contain specularities (local highlights) that are distributed throughout the frames.

Final representation: The video is represented as a set of key frames, and each key frame is represented by the visual words it contains and their position. This is the representation we use from here on for retrieval and data mining. The original raw images are not used other than for displaying the results.

3 Efficient Retrieval – Video Google

In this section we describe how the representation of section 2 can be used for object retrieval in videos, making an analogy with text based retrieval systems such as ‘Google’.



Fig. 2. Object query example II: searching for a Sony logo. First column: (top) Sony Discman image with the query region outlined in yellow and (bottom) close-up with detected elliptical regions superimposed. Second and third column: (top) frames from two different shots of ‘Groundhog Day’ with detected Sony logo outlined in yellow and (bottom) close-up of the image. The retrieved shots were ranked 1 and 4 (from 8 retrieved in total).

In text retrieval each document is represented by a vector of word frequencies – the ‘bag of words model’. Documents are then retrieved, in the first instance, by specifying a query as a set of words, and obtaining the documents corresponding to the vectors containing those words as components. It is usual to apply a weighting to the components of this vector [2], rather than use the frequency vector directly for indexing.

Here each key frame is represented by a weighted vector of the visual word frequencies it contains. An object query is specified by ‘lassoing’ the image object and thereby specifying its visual words and their configuration. This defines the query vector used for retrieval. The retrieved frames are ranked (in the first instance) according to the similarity (measured by angles) of their weighted vectors to this query vector.

Spatial consistency ranking: Up to this point we have simply used the ‘bag of (visual) words’ frequency representation, but we have not employed the spatial organization of the words. In a text search engine, such as Google, the ranking is increased for documents where the searched for words appear close together in the retrieved texts (measured by word order). This analogy is especially relevant for querying objects by a subpart of the image, where matched co-variant regions in the retrieved frames should have a similar spatial arrangement [11,12] to those of the outlined region in the query image. The idea is implemented here by re-ranking the retrieved frames based on a measure of spatial consistency. A search area is defined by the 15 nearest spatial neighbours (in the image) of each match, and each region which also matches within this area casts a vote for that frame. Matches with no support are rejected. The total number of votes determines the rank of the frame. More details of the method and other lessons borrowed from text retrieval [15] are given in [13].

Example queries: Figure 1 shows results of an object query for the movie ‘Groundhog Day’. The movie contains 5,640 keyframes (1 keyframe a second). Both the actual



Fig. 3. Mining Groundhog Day I. Examples of mined clusters at the 20 neighbourhood scale. Each row shows ten samples from one cluster. The first two rows correspond to faces of the two main characters. The next two rows show two different ties of the main character. The remaining rows show various objects that occur often in the movie. The images shown cover a rectangular convex hull of the matched configurations of viewpoint covariant regions within the frame plus a margin of 10 pixels. The rectangles are resized to squares for this display.

frames returned and their ranking are excellent – as far as it is possible to tell, no frames containing the object are missed (no false negatives), and the highly ranked frames all do contain the object (good precision). The query takes a fraction of a second on a 2GHz machine.



Fig. 4. Mining Groundhog Day II. Objects and scenes mined on the scale of (a) 50-neighbourhood and (b) 100-neighbourhood. The clusters extend over (a) 7,21,3 shots, (b) 7,3,5 shots (top-down).

Searching for objects from outside the movie: Figure 2 shows an example of searching for an object outside the ‘closed world’ of the film. The object (a Sony logo) is specified by a query image downloaded from the Internet. The image is preprocessed as outlined in section 2. Searching for images from other sources opens up the possibility for product placement queries, or searching movies for company logos, particular types of vehicles or buildings.

A demonstration version of Video Google is available at <http://www.robots.ox.ac.uk/~vgg/research/vgoogle/>.

4 Efficient Video Mining

In this section we describe how the representation of section 2 can be used for efficient mining of visual content from video.

The objective is to extract significant objects, characters and scenes in a video by determining their frequency of re-occurrence. An object is defined as a spatial configuration of vector quantized viewpoint co-variant regions – visual words. In analogy to the spatial consistency ranking used for retrieving images (described in section 3), a configuration consists of a visual word and its K spatial nearest neighbours. As the segmentation of an object is unknown in advance such a configuration is centred around every visual word in the movie and several scales are used (e.g. $K = 20, 50, 100$). The task then becomes that of measuring the re-occurrence of spatial configurations of visual words over an entire movie. Note the significant difference to the retrieval task described in section 3, where the segmentation of an object is given by a user outlining a query region and essentially only one object is searched for.

Measuring re-occurrence of configurations of visual words: The goal is to group configurations of visual words representing the same objects into clusters and count the number of occurrences within each cluster. A configuration re-occurs (is matched) if at least $M(= 3)$ of the visual words in the configuration are matched. Note that no



Fig. 5. Mining Casablanca. Examples of objects mined on the scale of 20-neighbourhood. Examples include the main characters, parts of clothing of other characters (e.g. uniforms) and objects (lamps) in Rick's bar.

geometric consistency on visual words (e.g. on their positions) in the configuration is required.

The algorithm consists of three stages. First, only configurations occurring in more than a minimum number of keyframes are considered for clustering. This filtering greatly reduces the data and allows us to focus on only significant (frequently occurring) configurations. Second, significant configurations are matched by a progressive clustering algorithm. At this stage, one object can be represented by multiple clusters which overlap spatially. This is because a configuration is placed at every visual word in a frame and therefore configurations are largely overlapping. Third, clusters representing one object are merged based both on spatial and temporal overlap in multiple keyframes. The algorithm is described in more detail in [14].

Examples: Figures 3 and 4 show samples from different clusters found for the scales of $K = 20, 50$ and 100 neighbourhood in the movie 'Groundhog Day'. Figure 5 shows samples from clusters found at the 20-neighbourhood scale in the movie 'Casablanca'. Generally, smaller consistent objects, e.g. faces and logos or objects which change background frequently or get partially occluded, tend to appear at the smaller scale. An ex-

ample would be the two clocks on the wall in the cafe (objects 7 and 8 of figure 3). On the larger scales we get (parts of) backgrounds, building fronts or the whole location. An interesting example is the ‘frames’ shop sign (object 9 of figure 3) which is extracted as a separate cluster at the 20-neighbourhood scale, and can be seen again as a subset of the a 100-neighbourhood scale cluster which covers the whole shop entrance (row 1 of figure 4b). Results on other videos and quantitative comparisons with ground truth are given in [14].

Acknowledgements. This research was supported by EC projects Vibes and CogViSys. We are grateful to the ViMining project of IMEDIA INRIA-Rocquencourt for travel funding.

References

1. A. Aner and J. R. Kender. Video summaries through mosaic-based shot and scene clustering. In *Proc. ECCV*. Springer-Verlag, 2002.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, ISBN: 020139829, 1999.
3. N. Boujemaa, J. Fauqueur, and V. Gouet. What’s beyond query by example? In *Trends and Advances in Content-Based Image and Video Retrieval*, 2004.
4. Y. Gong and X. Liu. Generating optimal video summaries. In *IEEE Intl. Conf. on Multimedia and Expo (III)*, pages 1559–1562, 2000.
5. D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, Sep 1999.
6. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC.*, pages 384–393, 2002.
7. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, 2001.
8. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*. Springer-Verlag, 2002.
9. S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. BMVC.*, pages 113–122, 2002.
10. F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proc. CVPR*, 2003.
11. F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or “How do I organize my holiday snaps?”. In *Proc. ECCV*, volume 1, pages 414–431. Springer-Verlag, 2002.
12. C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–534, May 1997.
13. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, Oct 2003.
14. J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *Proc. CVPR*, 2004.
15. D.M. Squire, W. Müller, H. Müller, and T. Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21:1193–1198, 2000.
16. B. Tseng, C.-Y. Lin, and J. R. Smith. Video personalization and summarization system. In *MMSP*, 2002.
17. T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. BMVC.*, pages 412–425, 2000.

Fast and Robust Short Video Clip Search for Copy Detection

Junsong Yuan^{1,2}, Ling-Yu Duan¹, Qi Tian¹, Surendra Ranganath², and Changsheng Xu¹

¹ Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{jyuan, lingyu, tian, xucs}@i2r.a-star.edu.sg

² Department of Electrical and Computer Engineering, National University of Singapore
elesr@nus.edu.sg

Abstract. Query by video clip (QVC) has attracted wide research interests in multimedia information retrieval. In general, QVC may include feature extraction, similarity measure, database organization, and search or query scheme. Towards an effective and efficient solution, diverse applications have different considerations and challenges on the abovementioned phases. In this paper, we firstly attempt to broadly categorize most existing QVC work into 3 levels: concept based video retrieval, video title identification, and video copy detection. This 3-level categorization is expected to explicitly identify typical applications, robust requirements, likely features, and main challenges existing between mature techniques and hard performance requirements. A brief survey is presented to concretize the QVC categorization. Under this categorization, in this paper we focus on the copy detection task, wherein the challenges are mainly due to the design of compact and robust low level features (i.e. an effective signature) and a kind of fast searching mechanism. In order to effectively and robustly characterize the video segments of variable lengths, we design a novel global visual feature (a fixed-size 144-d signature) combining the spatial-temporal and the color range information. Different from previous key frame based shot representation, the ambiguity of key frame selection and the difficulty of detecting gradual shot transition could be avoided. Experiments have shown the signature is also insensitive to color shifting and variations from video compression. As our feature can be extracted directly from MPEG compressed domain, lower computational cost is required. In terms of fast searching, we employ the active search algorithm. Combining the proposed signature and the active search, we have achieved an efficient and robust solution for video copy detection. For example, we can search for a short video clip among the 10.5 hours MPEG-1 video database in merely 2 seconds in the case of unknown query length, and in 0.011 second when fixing the query length as 10 seconds.

1 Introduction

As a kind of content-based video retrieval, Query by video clip (QVC) has posed many applications such as video copy detection, TV commercial & movie identification, and high level concept search. In order to implement a QVC solution, we have to solve the following challenges: 1) how to appropriately represent the video content and define similarity measure; 2) how to organize and access the very large dataset consisting of

large amounts of continuous video streams; and 3) the choice of a fast searching scheme to accelerate the query process. Towards an effective and efficient solution, diverse applications have different considerations and challenges on the abovementioned phases due to different search intentions. Different strategies and emphasis are thus applied. For example, the task of retrieving “similar” examples of the query at the concept level is associated with the challenge of capturing and modeling the semantic meaning inherent to the query [1] [2]. With an appropriate semantics modeling, those examples (a shot or a series of shots) with a similar concept as the query can be found. Here we are not concerned with search speed since the bottleneck against a promising performance is inherent to the gap between low-level perceptual features and high-level semantic concepts. In terms of video copy detection, an appropriate concept-level similarity measure is not required as the purpose is only to identify the presence or locate the re-occurrences of the query in a long video sequence. However, the prospective features or fingerprints are expected to be compact and insensitive to variations (e.g. different frame size, frame rate and color shifting) brought by digitization and coding. Particularly the search speed is a big concern. The reasons are twofold. Firstly, its application is usually oriented to a very large video corpus or a time-critical online environment; Secondly, the mostly used frame-based or window-based matching coupled with a shifting mechanism causes more serious granularity than the shot-based concept-level retrieval, wherein we have to quickly access much more high-dimensional feature points.

Based on the above discussions, we attempt to broadly categorize most existing QVC works into 3 levels, as illustrated in Fig.1. The production procedure of video content (left) is depicted and those associated QVC tasks at 3 different levels (right) are listed. Such categorization is expected to roughly identify common research issues, emphasis and challenges within different subsets of applications in diverse environments.

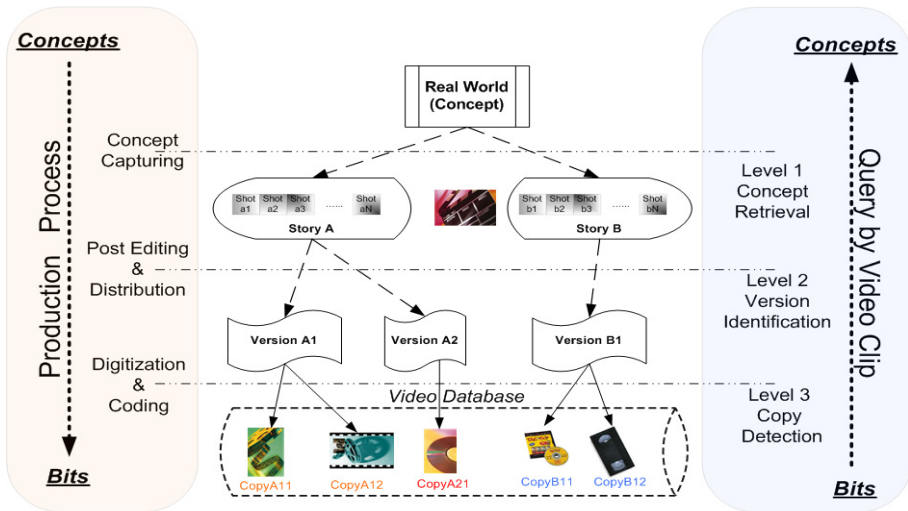


Fig. 1. A three layer framework for query by video clip.

Table 1. A Concretization of three-level QVC framework together with representative works.

Retrieval categories	Level 1 Video Copy Detection	Level 2 Video Title Identification	Level 3 Concept-based retrieval
Goals of retrieval operations	To retrieve copies of the query content, but variations with coding parameter changes are allowed, see below. (query and target instances share same content, same version, but may have coding variations) [15] [16][17][18][19][20][21][22][23][24] [25] [26] [27] [28] [29] [31] [32]	To retrieve contents with variations from the query content due to post editing. (query and target instances are from the same title, but are different versions. For instance, a movie and its trailers, an original advertisement and its reedited shortened version) [9] [10] [11][12] [13] [14] [20] [24] [28]	Concept search at semantic level. (query and target examples share "similar" concept or content, for instance, to retrieve video clips of a Boeing 747's take off scene, to find out all the free-throw scenes in a basketball match) [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [19]
Variations allowed	Coding variations, including <ul style="list-style-type: none"> • Frame rate • Bit rate • Frame size • Color shift • Digitization noises etc. 	Shot level editing (shot removal, insertion, summary etc.); Image frame editing (caption insertion, Logo insertion, frame shift, etc.); same movie but in different language editions, etc.	Many instances of the same concepts, but appearing in different setting and environments, including but not limited to camera angle variation, scale, acquisition time and place variations, etc.
Likely features used	Mainly low level features such as color, gray level features, independent of video structure	Low level features and features from image, shot or object levels; content modeling	Refined low level features; Object level features and high level concept features; Graph model, motion model, etc.
Research challenges	Precise similarity measure at concept-level is not required, but the signatures need to be compact and robust against various coding variations; [15] [16] [19] [20] [22] [24] [25] [26] The target database usually contains unlabeled video streams or are considering on-line application environment, therefore efficient feature extraction and fast search is required; [15] [17] [18] [20] [21] [22] [23] [27] [29] [31] [32] Database should be organized properly to support query with various length	Efficient detection and robust representation against various post-editing effects are needed; [10] [11] [12] [13] [14] [20] [24] [28]	Robust detection of high level concepts, such as in-door, out-door (TRECVID concepts); [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]; Effective organization and indexing and the large dataset to support fast access and efficient query [30]
Existing Works	NTT [17] [18] [32], I ² R [25] [26], RMIT [22], University of Kansas [21], IBM [16], Philips [29], Intel [15], NEC [23], INRIA [20]	Berkeley University [11] [12], NUS [13], Columbia University [10], Oakland University [14], INRIA [20], Fraunhofer Institute [24]	TRECVID [1], Michigan State University [3], Columbia University [9] [10], I ² R [6] [7] [8] etc.

Under this framework, our work in this paper is focused on video copy detection, the lowest search level (See Sections 3, 4, 5). We want to jointly take into account the robustness issue and the search speed issue to complete the efficient and effective detection. The experimental dataset includes 10.5 hours video collections and in total 84 given queries with the length ranging from 5 to 60 seconds are performed. Our experiments have shown that both fast search speed and good performance can be accomplished at the lowest retrieval level.

2 Related Works

After a comprehensive literature review [1-32], we concretize the framework as listed in the Table 1. The references are roughly grouped around application intentions and their

addressed research challenges respectively. Due to limited space, no detailed comparison will be given here.

3 Feature Extraction for Video Copy Detection

In video copy detection, the signature is required to be compact and efficient with respect to large database. Besides, the signature is also desired to be robust to various coding variations mentioned in Table 1. In order to achieve this goal, many signature and feature extraction methods are presented for the video identification and copy detection tasks [11] [15] [16] [26] [28] [29].

As one of the common visual features, color histogram is extensively used in video retrieval and identification [15] [11]. [15] applies compressed domain color features to form compact signature for fast video search. In [11], each individual frame is represented by four 178-bin color histograms in the HSV color space. Spatial information is incorporated by partitioning the image into four quadrants. Despite certain level of success in [15] and [11], the drawback is also obvious, e.g. color histogram is fragile to color distortion and it is inefficient to describe each individual key frame using a color histogram as in [15].

Another type of feature which is robust to color distortion is the ordinal feature. Hampapur et al. [16] compared performance of using ordinal feature, motion feature and color feature respectively for video sequence matching. It was concluded that ordinal signature had the best performance. The robustness of ordinal feature was also proved in [26]. However, based on our experiments, we believe better performance could be achieved by combining ordinal features and color range features appropriately, with the former providing spatial information and the latter providing range information. Experiments in Section 5 support these conclusions. As a matter of fact, many works such as [3] and [14] also incorporate the combined feature in order to improve the performance of retrieval and identification.

Generally, the selection of ordinal feature and color feature as signature for copy detection task is motivated by the following reasons:

- (1) Compared with computational cost features such as edges, texture or refined color histograms which also contain spatial information (e.g. *color coherent vector* applied in [28]), they are inexpensive to acquire
- (2) Such features can form compact signatures [29] and retain perceptual meaning
- (3) Ordinal features are immune to global changes in the quality of the video and also contain spatial information, hence are a good complement to color features [26]

3.1 Ordinal Feature Description

In our approach, we apply Ordinal Pattern Distribution (*OPD*) histogram proposed in [26] as the ordinal feature. Different from [26], the feature size is further compressed in this paper, by using more compact representation of I frames. Figure 2 depicts the operations of extracting such features from a group of frames.

For each channel $c = Y, Cb, Cr$, the video clip is represented by *OPD* histograms as:

$$H_c^{OPD} = (h_1, h_2, \dots, h_l, \dots, h_N) \quad 0 \leq h_i \leq 1 \quad \text{and} \quad \sum_i h_i = 1 \quad (1)$$

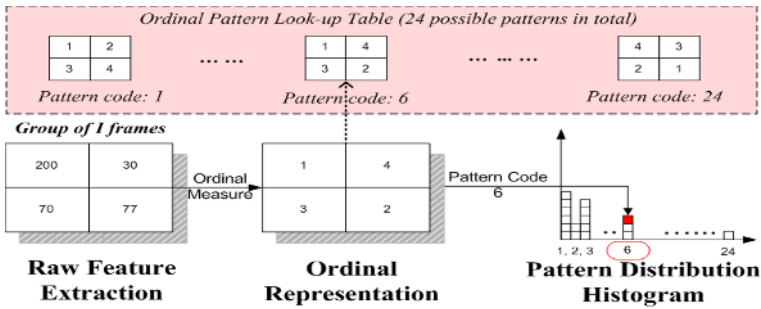


Fig. 2. Ordinal Pattern Distribution (OPD) Histogram.

Here $N=4! = 24$ is the dimension of the histogram, namely the number of possible patterns mentioned above. The total dimension of the ordinal feature is $3 \times 24=72$.

The advantages of using *OPD* histograms as visual features are two fold. First, they are robust to frame size change and color shifting as mentioned above. And secondly, the contour of the pattern distribution histogram can describe the whole clip globally; therefore it is insensitive to video frame rate change and other local frame changes compared with key frame representation.

3.2 Color Feature

For the color feature, we characterize the color information of a GoF by using the cumulative color information of all the sub-sampled I frames in it. For computational simplicity, Cumulative Color Distribution (*CCD*) is also estimated using the DC coefficients from the I frames.

The cumulative histograms of each channel ($c=Y, Cb, Cr$) can be defined as:

$$H_c^{CCD} = \frac{1}{M} \sum_{i=b_k}^{b_{k+M}-1} H_i(j) \quad j = 1, \dots, B \quad (2)$$

where H_i denotes the color histogram describing an individual I frame in the segment. M is the total number of I frames in the window and B is the color bin number. In this paper, $B = 24$ (uniform quantization). Hence, the total dimension of the color feature is also $3 \times 24=72$, representing three color channels.

4 Similarity Search and Copy Detection

For visual signature matching, Euclidean distance $D(\cdot, \cdot)$ is used to measure distance between the query Q (represented by H_Q^{OPD} and H_Q^{CCD} , both are 72-d signatures) and the sliding matching window SW (represented by H_{SW}^{OPD} and H_{SW}^{CCD} , both are 72-d signatures). The integrated similarity S is defined as the reciprocal of linear combination

of the *average* distance of *OPD* histograms and the *minimum* distance of *CCD* histograms in the Y, Cb, and Cr channels:

$$D^{OPD}(H_Q^{OPD}, H_{SW}^{OPD}) = \frac{1}{3} \sum_{c=Y,Cb,Cr} D(H_Q^{OPD}, H_{SW}^{OPD}) \quad (3)$$

$$D^{CCD}(H_Q^{CCD}, H_{SW}^{CCD}) = \underset{c=Y,Cb,Cr}{Min} \{D(H_Q^{CCD}, H_{SW}^{CCD})\} \quad (4)$$

$$S(H_Q, H_{SW}) = \frac{1}{w \times D^{OPD} + (1 - w) \times D^{CCD}} \quad (5)$$

Let the similarity metric array be $\{S_i; 1 \leq i \leq m + n - 1\}$ corresponding to similarity values of $m + n - 1$ sliding windows, where n and m are the I frame number of the query clip and the target stream respectively. Based on [17] and [32], the search process can be accelerated by skipping unnecessary steps. The number of skipped steps w_i is given as:

$$w_i = \begin{cases} \lceil \sqrt{2}D(\frac{1}{S_i} - \theta) \rceil + 1 & \text{if } S_i < \frac{1}{\theta} \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

where D is the number of I frames of the corresponding matching window. θ is the predefined skip threshold.

After the search, potential start position of the match is determined by a local maximum above the threshold, which fulfills the following conditions:

$$S_{k-1} \leq S_k \geq S_{k+1} \text{ and } S_k > \max\{T, m + k\sigma\} \quad (7)$$

where T is the pre-defined preliminary threshold, m is the mean and σ is the deviation of the similarity curve; k is an empirically determined constant. Only when similarity value satisfies (7), is it treated as the detected instance. In our experiments, w in (5) is set to 0.5, and θ in (6) is set to 0.05, and T in (7) is set to 6.

5 Experimental Results

All the simulations were performed on a P4 2.53G Hz PC (512 M memory). The algorithm was implemented in C++. The query collection consists of 83 individual commercials which varied in length from 5 to 60 seconds and one 10-second long news program lead-out clip (Fig. 3). All the 84 given clips were taken from ABC TV news programs. The experiment sought to identify and locate these clips inside the target video collection, which contains 22 streams of half-hour broadcast ABC news video (obtained from TRECVID news dataset [1]). The 83 commercials appear in 209 instances in these half-hour news programs; and the lead-out clip appears in total 11 instances. The re-occurrence instances usually have color shifting, I frame shifting and frame size variations with respect to the original query. All the video data were encoded in MPEG1 at 1.5 Mb/sec with image size of 352×240 or 352×264 and frame rate of 29.97 fps. It is compressed with the frame pattern IBBPBBPBBPBB, with I frame temporal resolution around 400 ms. Fig. 3 and Fig. 4 give two examples of extracted features.

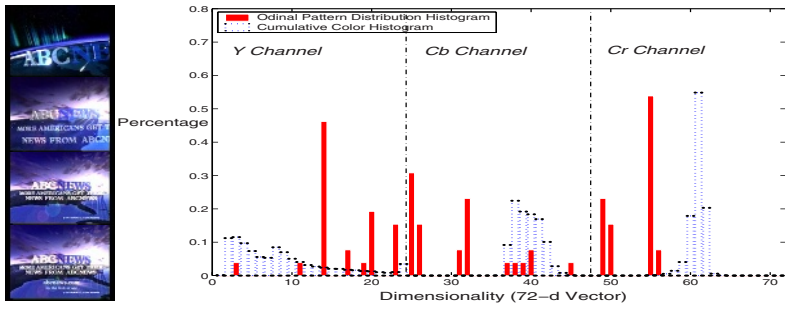


Fig. 3. ABC News program lead-out clip (left, 10 sec) and its *CCD* and *OPD* signatures (right).

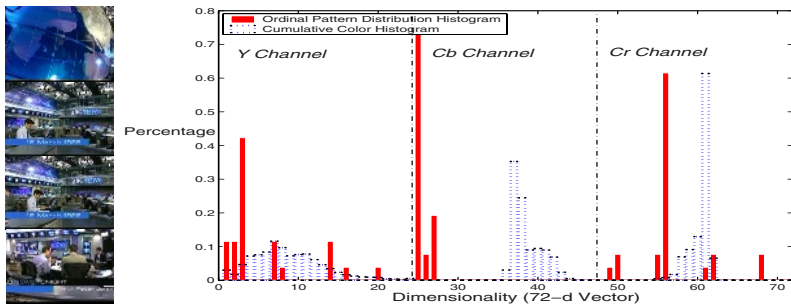


Fig. 4. ABC News program lead-in clip (left, 10 sec) and its *CCD* and *OPD* signatures (right).

We note here that the identification and retrieval of such repeated *non-news* sections inside a video stream helps to reveal the video structure. These sections include TV commercials, program lead-in/lead-out and other Video Structure Elements (*VSE*) which appear very often in many types of video to indicate starting or ending points of a particular video program, for instance, news programs or replay of sports video.

Table 2 gives the approximate computational cost of the algorithm. The task is to search for instances of the 10 second long lead-out clip (Fig. 3) in the 10.5 hour MPEG-1 video dataset. The *Feature Extraction* step includes DC coefficient extraction from the compressed domain, the formation of color histogram ($3 \times 24-d$) of each I frame (H_i histogram in (2)). This step could be done off-line for the 10.5-hour database. On the other hand, *Signature Processing* consists of the procedures to form *OPD* and *CCD* signatures for the specific matching windows during the active search. Therefore its cost may vary according to the length of the window, namely the length of the query. If the query length is known or fixed beforehand, signature processing step could also be done off-line. In that case, the only cost of active search is *Similarity Calculation*. In our experiment, similarity calculation through a video database of 10.5 hours needs only 11 milliseconds.

The performance of searching for the instances of the given 84 clips in the 10.5 hour video collection is presented in Fig. 5. From the experimental results we found that a

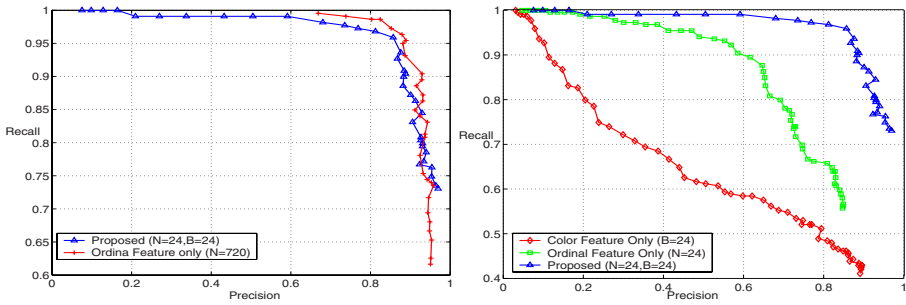


Fig. 5. Performance comparison using different feature: proposed features vs. 3×720 -d *OPD* feature (left); proposed features vs. 3×24 -d *CCD* feature and 3×24 -d *OPD* feature respectively (right); the detection curves are generated by varying the parameter k in (7) ($Precision = \text{detects} / (\text{detects} + \text{false alarms})$) ($Recall = \text{detects} / (\text{detects} + \text{miss detects})$).

large part of the false alarms and missed detections are mainly caused by the I frame *shifted matching* problem, when the sub-sampled I frames of the given clip and that of the matching window are not well aligned in temporal axis. Although the matching did not yield 100% accuracy using the proposed signatures (72-d *OPD* and 72-d *CCD*), it still obtains performance which is comparable with that of [26], where only *OPD* with $N=720$ is considered. However, compared with [26] whose feature size is $3 \times 720=2160$ dimension, our proposed feature is as small as a $(3 \times 24+3 \times 24) = 144$ dimensional vector, 15 times smaller than that of [26]. Besides, in terms of Fig. 5, it is obvious that better performance can be achieved by using the combined features than using only *CCD* (color feature) or only *OPD* (ordinal feature) respectively.

6 Conclusion and Future Work

In this paper, we have presented a three-level QVC framework in terms of how to differentiate the diverse “similar” query requests. Although huge amounts of QVC research have been targeted in different aspects (e.g. feature extraction, similarity definition, fast search scheme and database organization), few work has tried to propose such a framework to explicitly identify different requirements and challenges based on rich applications. A closely related work [28] has just tried to differentiate the meanings of “similar” at different temporal levels (i.e. frame, shot, scene and video) and discussed various strategies at those levels. According to our experimental observation and comparisons among different applications, we believe that a better interpretation of the term of

Table 2. Approximate Computational Cost Table (CPU time).

Task: Search for a 10 second long query clip	Feature Extraction	Active Search		
		Signature Processing		Similarity Calculation of Signatures
		Ordinal Feature	Color Feature	
10.5 h MPEG1 Video (93,725.1 frames)	1178.034 sec	0.969 sec	0.688 sec	0.011 sec

“similar” is inherent to the user-oriented intentions. For example, in some circumstances, the retrieval of “similar” instances is to detect the exact duplicate or re-occurrences of the query clip. Sometimes, the “similar” instances may designate the re-edited versions of the original query. Besides, searching “similar” instances could also be the task of finding video segments sharing the same concept or having the same semantic meaning as that of the query. Different bottlenecks and emphasis exist at these different levels.

Under the framework, we have provided an efficient and effective solution for video copy detection. Instead of the key frames-based video content representation, the proposed method treats the video segment as a whole, which is able to handle video clips of variable length (e.g. a sub-shot, a shot, or a group of shots). However, it does not require any explicit and exact shot boundary detection.

The proposed *OPD* histogram has experimentally proved to be a useful complement to the *CCD* descriptor. Such an ordinal feature can also reflect a global distribution within a video segment by the accumulation of multiple frames. However, the temporal order of frames within a video sequence has not yet been exploited sufficiently in *OPD*, and also in *CCD*. Although our signatures are useful for those applications irrespective of different shot order (such as the commercial detection in [13]), the lack of frame ordering information may make the signatures less distinguishable. Our future work may include how to incorporate temporal information, how to represent the video content more robustly and how to further speed up the search process.

References

- [1] <http://www-nlpir.nist.gov/projects/trecvid/>. *Web site*, 2004
- [2] N.Sebe et al., “The state of the art in image and video retrieval,” In *Proc. of CIVR’03*, 2003
- [3] A. K. Jain et al., “Query by video clip,” In *Multimedia System*, Vol. 7, pp. 369-384, 1999
- [4] D. DeMenthon et al., “Video retrieval using spatio-temporal descriptors,” In *Proc. of ACM Multimedia’03*, pp. 508-517, 2003
- [5] Chuan-Yu Cho et al., “Efficient motion-vector-based video search using query by clip,” In *Proc. of ICME’04*, Taiwan, 2004
- [6] Ling-Yu Duan et al., “A unified framework for semantic shot classification in sports video,” To appear in *IEEE Transaction on Multimedia*, 2004
- [7] Ling-Yu Duan et al., “Mean shift based video segment representation and applications to replay detection,” In *Proc. of ICASSP’04*, pp. 709-712, 2004
- [8] Ling-Yu Duan et al., “A Mid-level Representation Framework for Semantic Sports Video Analysis,” In *Proc. of ACM Multimedia’03*, pp. 33-44, 2003
- [9] Dong-Qing Zhang et al., “Detection image near-duplicate by stochastic attribute relational graph matching with learning,” in *Proc. of ACM Multimedia’04*, New York, Oct. 2004
- [10] Alejandro Jaimes, Shih-Fu Chang and Alexander C. Loui, “Detection of non-identical duplicate consumer photographs,” In *Proc. of PCM’03*, Singapore, 2003
- [11] S. Cheung and A. Zakhor, “Efficient video similarity measurement with video signature,” In *IEEE Trans. on Circuits and System for Video Technology*, vol. 13, pp. 59-74, 2003
- [12] S.-C. Cheung and A. Zakhor, “Fast similarity search and clustering of video sequences on the world-wide-web,” To appear in *IEEE Transactions on Multimedia*, 2004.
- [13] L. Chen and T.S. Chua, “A match and tiling approach to content-based video retrieval,” In *Proc. of ICME’01*, pp. 301-304, 2001
- [14] V. Kulesh et al., “Video clip recognition using joint audio-visual processing model,” In *Proc. of ICPR’02*, vol. 1, pp. 500-503, 2002

- [15] M.R. Naphade et al., "A Novel Scheme for Fast and Efficient Video Sequence Matching Using Compact Signatures," In *Proc. SPIE, Storage and Retrieval for Media Databases 2000*, Vol. 3972, pp. 564-572, 2000
- [16] A. Hampapur, K. Hyun, and R. Bolle., "Comparison of Sequence Matching Techniques for Video Copy Detection," In *SPIE. Storage and Retrieval for Media Databases 2002*, vol. 4676, pp. 194-201, San Jose, CA, USA, Jan. 2002.
- [17] K. Kashino et al., "A Quick Search Method for Audio and Video Signals Based on Histogram Pruning," In *IEEE Trans. on Multimedia*, Vol. 5, No. 3, pp. 348-357, Sep. 2003
- [18] K. Kashino et al., "A quick video search method based on local and global feature clustering," In *Proc. of ICPR'04*, Cambridge, UK, Aug. 2004
- [19] A.M. Ferman et al., "Robust color histogram descriptors for video segment retrieval and identification," In *IEEE Trans. on Image Processing*, vol. 1, Issue 5, May 2002
- [20] Alexis Joly, Carl Frelicot and Olivier Buisson, "Robust content-based video copy identification in a large reference database," In *Proc. of CIVR'03, LNCS 2728*, pp. 414-424, 2003
- [21] Kok Meng Pua et al., "Real time repeated video sequence identification," In *Journal of Computer Vision and Image Understanding*, vol. 93, pp. 310-327, 2004
- [22] Timothy C. Hoad, et al., "Fast video matching with signature alignment," In *SIGIR Multimedia Information Retrieval Workshop 2003 (MIR'03)*, pp. 263-269, Toronto, 2003
- [23] Eiji Kasutani et al., "An adaptive feature comparison method for real-time video identification," In *Proc. of ICIP'03*, 2003
- [24] Nicholas Diakopoulos et al., "Temporally Tolerant Video Matching," In *SIGIR Multimedia Information Retrieval Workshop 2003 (MIR'03)*, Toronto, Canada, Aug. 2003
- [25] Junsong Yuan et al. "Fast and Robust Short Video Clip Search Using an Index Structure," in *ACM Multimedia Workshop on Multimedia Information Retrieval (MIR'04)*, 2004
- [26] Junsong Yuan et al., "Fast and Robust Search Method for Short Video Clips from Large Video Collection," in *Proc. of ICPR'04*, Cambridge, UK, Aug. 2004
- [27] Sang Hyun Kim and Rae-Hong Park, "An efficient algorithm for video sequence matching using the modified Hausdorff distance and the directed divergence," in *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 12 pp. 592-596, July 2002
- [28] R. Lienhart et al., "VisualGREP: A Systematic method to compare and retrieve video sequences," In *SPIE. Storage and Retrieval for Image and Video Database VI*, Vol. 3312, 1998
- [29] J. Oostveen et al., "Feature extraction and a database strategy for video fingerprinting," In *Visual 2002, LNCS 2314*, pp. 117-128, 2002
- [30] Jianping Fan et al., "Classview: hierarchical video shot classification, indexing and accessing," In *IEEE Trans. on Multimedia*, Vol. 6, No. 1, Feb. 2004
- [31] Chu-Hong Hoi et al., "A novel scheme for video similarity detection," In *Proc. of CIVR'03, LNCS 2728*, pp. 373-382, 2003
- [32] Akisato Kimura et al., "A Quick Search Method for Multimedia Signals Using Feature Compression Based on Piecewise Linear Maps," In *Proc. of ICASSP'02*, 2002

Mining Large-Scale Broadcast Video Archives Towards Inter-video Structuring

Norio Katayama¹, Hiroshi Mo¹, Ichiro Ide², and Shin'ichi Satoh¹

¹ National Institute of Informatics

2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

{katayama, mo, satoh}@nii.ac.jp

² Graduate School of Information Science, Nagoya University

1 Furo-cho, Chikusa-ku, Nagoya-shi, Aichi 464-8601, Japan

ide@is.nagoya-u.ac.jp

Abstract. The current computer technology enables us to build huge broadcast video archives which had been a future dream. Today, even the hard disk recorders on the market are capable of recording several hundred hours of broadcast video. It is naturally perceived that a huge amount of broadcast video would be a useful corpus for multimedia indexing and mining research. Based on this viewpoint, we designed and constructed a broadcast video archive system having sufficient capacity and functionality to serve as the testbed for indexing and mining research. The system can capture multiple channels (currently seven channels) all-day broadcast video streams simultaneously, up to 6000 hours, and program-specific broadcasts, currently a news program for more than three years so far. This paper discusses design and implementation issues of the video archive system and then introduces our research efforts utilizing the archives as huge multimedia corpora.

1 Introduction

The current computer technology enables us to build huge broadcast video archives which had been a future dream. Today, even the hard disk recorders on the market are capable of recording several hundred hours of broadcast video. It is naturally perceived that a huge amount of broadcast video would be a useful corpus for multimedia indexing and mining research. In addition, we should note that the speed of the technology progress is considerably fast. We observe that the capacity of hard disk drives grows ten times in every three years. Thus it might be possible, in five to ten years, to realize a personal set top box (STB) with terabytes to petabytes of disk drives, which can store thousands to millions of hours of videos. In such circumstances, it is crucial to establish component technologies of indexing and mining that are applicable to such huge multimedia corpora.

Based on this viewpoint, we designed and constructed a broadcast video archive system having sufficient capacity and functionality to serve as a testbed for multimedia indexing and multimedia mining research, which reflects a realistic scale for the future STB. In order to reflect the important nature of the huge broadcast video corpus, the system captures multiple channels simultaneously, 24 hours a day, as MPEG video files. The system also captures related text information including closed-caption text and electronic program guide (EPG) information. In realizing the system, instead of special

hardware, we employ commodities for the components such as UNIX workstations, RAID disk arrays, and MPEG capture cards and closed-caption decoder cards installed in PCs. Captured data are managed by Oracle DBMS, and the system provides uniform view of access to the data for clients via Java JDBC API. We also developed the experimental video browser system which is intended to be used as the software platform of the system enabling rapid prototyping of video applications and video analysis software. This paper discusses design and implementation issues of the video archive system and then introduces our research efforts utilizing the archives as huge multimedia corpora.

2 Broadcast Video Archive System for Multimedia Research

2.1 Design Issues

In order to reflect the large-scale and dynamic nature of the broadcast video streams, the video archive system should meet several design issues. We first discuss desired functions of the system.

Desired video archives suited to multimedia indexing and mining research may need to keep capturing videos as a long period as possible, i.e., several years, while at the same time, they are required to capture as many streams (or channels) as possible for 24 hours a day. However, it is impossible to satisfy both requests at the same time, since in order to do this, required volume size of storage easily becomes prohibitively large. For important directions of multimedia indexing and mining research, there are mainly three demands for video archives making a compromise between the above two requests:

Diversity: Capture of all-day video stream of as many channels as possible regardless of types of programs for a relatively small period, e.g., several days to several weeks.

Continuity: Capture of particular programs broadcasted daily or weekly for a long period, e.g., several months to several years.

Autonomy: Dynamic registration of captured data to archives reflecting everyday broadcasts.

Research efforts to handle continuous video streams (e.g., [1] realizes browsing of 24-hour broadcast videos) require diverse video archives. On the other hands, research attempts to analyze in-depth contents of specific types of programs (e.g., [2,3,4,5] concentrate on news, [6] on sports, [7] on cooking videos, etc.) may need continuous archives of specific types of programs. Therefore, desired video archives should satisfy both diversity and continuity. In realizing diverse and continuous video archives, daily or even hourly capture, registration, and management of video files are required. It obviously is impossible to manually construct the video archives having these characteristics. Automated capture, registration, and management are indispensable. Thus desired video archives may need to satisfy autonomy also, which tends to be lacking in the former attempts. CMU Informedia project [8] built the video archive system which satisfies autonomy for specific types of programs, namely, news and documentaries. But it does not meet diversity. We seek for all three demands at the same time for the video archive system.

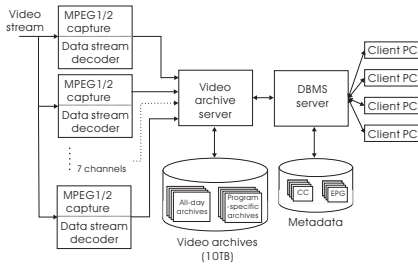


Fig. 1. Block Diagram of the Archive System

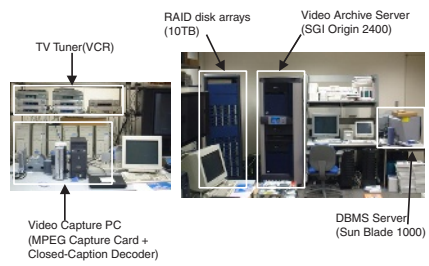


Fig. 2. Overview of the System

2.2 Implementation Issues

Recently, many commodities are becoming available which can be used as key components of the video archive system. However, there are no product which sufficiently satisfies our requirements. In implementing the video archive system, we decided to use commodities as components such as RAID disk arrays, MPEG capture cards, closed caption decoders, etc., and combine them on our own. Even though they are available as commodities, it is required to connect all these components in an integrated way, which is not trivial task, but rather challenging. In this section, we discuss implementation issues on the system by combining commodities.

Figure 1 shows the block diagram of the broadcast video archive system taking the demands discussed in the previous section into account. We use broadcast streams from seven channels (all terrestrial broadcasts available in Tokyo area) as sources of the video streams. The system can capture seven video streams simultaneously, and generate all-day video streams into one hour long MPEG files for each channel. In addition to video streams, the system also captures related text information including closed-caption (CC) text from data broadcasts, and electronic program guide (EPG) information from web. The system has an MPEG capture card (Canopus MVR-D2000) and a closed-caption decoder card (Systec Moji-Vision 550) installed in a PC for each channel, in total seven channels, to obtain required data.

Captured data are fed to the video archive server with RAID disk arrays. As the video archive server, SGI Origin 2400 is used, having 10TB RAID disk arrays connected through fiber channels. PCs are controlled by the server regularly, and to enable remote procedure calls (RPC) from the UNIX server to Windows PC, we use the Java remote method invocation (RMI) mechanism. The 10TB storage is split into 7TB for all-day streams (for diverse archives), and 3TB for particular programs (for continuous archives). Due to disk capacity limitations, the system keeps recent one month archives, about 6000 hours in MPEG-1 format in the 7TB storage. At the same time, the system captures one news program from NHK (the largest broadcast station in Japan) everyday, and keeps them persistently, so far for more than three years. For this type of archives, we capture video streams into MPEG-2 format to guarantee higher quality for the purpose of high-precision video processing such as face detection and motion analysis. The system captures these video streams in fully automated way, so it also satisfies autonomy. The system is also easily reconfigurable to capture other programs in addition to news.

Another component of the system is the DBMS server. We use Sun Blade 1000 for the hardware platform and Oracle for the DBMS software. The DBMS server regularly checks if there are any change on the video archives, and it autonomously reflects their changes to the database. The current implementation stores the following metadata in the DBMS server.

- Properties of MPEG video files (file path, broadcasted time, duration, channel number, etc.)
- Closed caption texts (each piece of closed caption text, broadcasted time, channel number, etc.)
- Electronic program guide (EPG) information (program guide text, start time of each program, channel number of the program, etc.)

Closed caption texts and EPG information are indexed by Oracle InterMedia Text so that they can be retrieved by the full-text search. These metadata provide the fundamental methods for locating and retrieving a particular video segment from the huge “video stream space”.

By implementation techniques described here, the system is realized using commodities as components, and it works well satisfying requirements discussed in the previous section in an integrated way.

2.3 Software Platform for Prototyping

This video archive system is intended to be used by researchers for indexing and mining research. In principle, they can use this system by accessing MPEG files stored in the RAID disk arrays and their metadata in the Oracle database. However, it is not easy to develop research prototypes or experimental computer programs from scratch. Therefore, it is crucial to implement a software platform which provides handy application programming interface (API) for indexing and mining research. The software platform reduces not only the development cost but also the maintenance cost of the video archive system. As is common with other applications, the API encapsulates the internal configuration of the system and enables the system maintainer to evolve the configuration without interfering the end users of the system.

On designing the software platform, we focused on the following three requirements:

- (1) Researchers should be able to develop their own computer programs easily which use the video files and their metadata stored in the system.
- (2) The software platform should enable researchers to develop their programs on their own machines. This means that the software platform should provide the remote access to the video archive system.
- (3) The software platform should work on various hardware platforms since the video archive system itself is an integration of various hardware components and since the researchers use a wide variety of hardware platforms including UNIX workstations and Windows PCs.

In order to fulfill these requirements, we employ Java programming language. Java is available on a wide variety of hardware platforms and it has a standard API for accessing

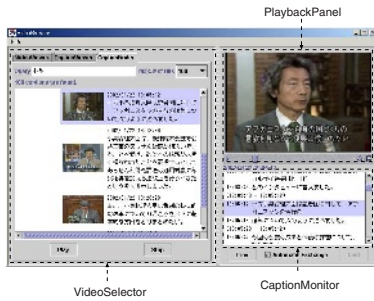


Fig. 3. Video Browser

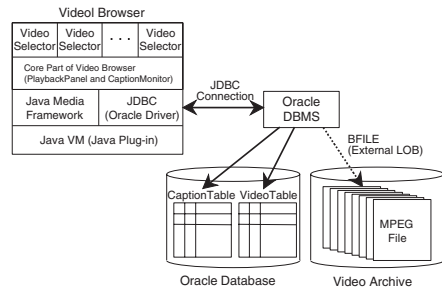


Fig. 4. Block Diagram of Video Browser

remote DBMSs which is called JDBC (Java Database Connectivity). In addition, the combination of Java and the relational DBMS is so popular that we can take advantage of various information resources and development tools on the Web and on the market.

The API determines the usability of the software platform. In order to promote the rapid prototyping by researchers, we developed a Java applet which plays the role of the basis for prototyping. The applet is called “video browser” and has the following capabilities:

- Provides the fundamental GUI components for prototype systems.
- Enables researchers to accommodate their custom GUI components with providing the fundamental API for accessing the video archive.

The video browser consists of three types of components: PlaybackPanel, CaptionMonitor, and one or more VideoSelectors (Figure 3 and 4). PlaybackPanel plays a video stream, while CaptionMonitor displays text information synchronously to the video playback. VideoSelectors are GUI components for browsing/searching the video archive. Since they are placed on the tabbed pane, users can use multiple VideoSelectors interchangeably. By default, the video browser contains three VideoSelectors: VideoViewer (directory listing of video files), CaptionViewer (directory listing of textual information, i.e., closed caption and EPG), and CaptionSearch (full-text search of closed caption and EPG). In addition to these VideoSelectors, researchers can add their own VideoSelectors for testing their original indexing and mining methods.

As shown in Figure 4, the video browser accesses the MPEG files in the video archive through Oracle DBMS. Since Oracle DBMS has the mechanism to access external binary large objects, the video browser obtains both metadata and MPEG files only from JDBC connection. The obtained MPEG stream is fed to the player of Java Media Framework. Thus, the video browser requires only the JDBC connection to the DBMS server; it does not need other file access methods, such as network file system. This expands the coverage of the remote access to the video archive system.

3 Research Challenges in Mining Broadcast Video Corpora

We are now conducting multimedia indexing and mining research by taking advantage of the video archive system mentioned above. In this section, we introduce our recent efforts with presenting some lessons learned from using huge broadcast video corpora.

3.1 Inter-video Structuring with Associating Video Segments

Broadcast video corpora potentially contains wide variety of information, which might be useful for developing advanced multimedia applications and evolving multimedia technology. While there exist various approaches for mining video corpora, one of the most fundamental approach is associating video segments to determine the inter-video structure among video streams. Association may cover any type of relationships, e.g., the appearance of the same person, recorded at the same location, dealing the same topic, etc. For example, in [9], broadcast videos of a daily news program are divided into video segments based on the topic boundary, and then the resultant segments are mutually associated based on the relevance between topics. By this approach, video segments are interweaved with the threads of topics. As mentioned above, our archive accumulates a daily news program for more than three years. Finding the topical relationship in a video corpus provides effective clues for inter-video structuring. Although the detection of inter-video relationships is the first step toward inter-video structuring, we believe that it is the essential and indispensable foothold.

3.2 Multimodality: Integrating Multimodal Information

Needless to say, multimodality is one of the most important property of broadcast video corpora. Video streams contains, textual(open and closed caption text), audio, and visual information. Since each type of information has different nature, it is important to integrate advantages of each type of information. For example, in [3], the face-name association is obtained based on the cooccurrence between a face image and a person name text. In [10], the key frame for the specific topic is detected based on the cooccurrence between frame images and topic threads. These methods successfully detect video-text relationships by integrating multimodal information. In [11], textual and visual information are used interchangeably in the video retrieval process. Since textual information is effective in expressing topic keywords and object names (persons, locations, etc.), it is used as a strongly restrictive filter. On the other hand, visual information is used for seeking useful or interesting video scenes from a bunch of video segments through browsing. This method reflects pros and cons of textual and visual information. The advantage of textual information is small processing cost and strong expressive power in restricting topics. On the other hand, the advantage of visual information is the efficiency in human perception, i.e., you can rapidly perceive the contents at a glance. As illustrated by above examples, the integration of multimodal information is an important strategy for utilizing huge multimedia corpora.

3.3 Mining Rare but Strong Inter-video Relationships

Compared with textual information, it is quite difficult to use visual information for the clues to inter-video structuring. This is mainly because visual information is more subject to ambiguity due to the diversity of object appearances caused by the difference in pause, composition, lighting conditions, etc. However, if we focus on some particular type of visual information, it is possible to obtain strong inter-video relationships as mentioned below. Since this type of relationships can be obtained only for some specific type of visual information, they may rarely exist. However, when we have huge multimedia corpora, strong relationships are invaluable clues to inter-video structuring.

In [12], identical video segments are detected in a news program series, and it is reported that identical video segments are sometimes used for presenting particular topics or objects. In general, the frequency of identical video segments is not so high, but some symbolical and topical shots are often used repeatedly when some particular topic attracts public attention. The occurrence of identical shots are rather rare but once they are detected, they provide strong inter-video relationships. This strategy may not work with a small set of broadcast videos but it can be a strong tool for inter-video structuring when it is used with a large-scale broadcast video corpus.

Another example of mining rare relationships is face sequence matching with finding closest pairs[13]. Finding the similar face sequences is a basic function for finding the appearance of the same person. One of the difficulty in finding the similar face sequences is the diversity of recording conditions, e.g., lighting, pause, facial expression, etc. Although the conditions may vary for each face sequence, two face sequences may happen to contain two face image pair that are very close to each other. Based on this viewpoint, we developed a face sequence matching method that evaluates the similarity of face sequences by the similarity between the closest pair. Although the search for the closest pair is computation intensive, this method gives better precision than the method which evaluates the similarity of face sequences by comparing best-frontal face image pairs.

3.4 Online Management of Large-Scale Video Corpora

From the implementation aspect, an important nature of the broadcast video archive system is that it is an online system and archives grow continuously. In order to reflect the latest broadcast information, indexing and mining methods must be adapted to the online processing. In addition, as the storage amount and the number of recording channels increase, the requirement for the online processing should be more crucial. At this moment, we do not have concrete solution to this problem but we expect that the database technology would play an important role in extending the scalability.

4 Conclusions

We designed and constructed a broadcast video archive system having sufficient capacity and functionality to serve as a testbed for multimedia indexing and multimedia mining research. Three demands were pointed out for desired video archive system, i.e., diversity,

continuity, and autonomy. Actual implementation of the system satisfying these demands is shown using commodities as its key components, such as UNIX workstations, RAID disk arrays, and MPEG capture cards and closed-caption decoder cards installed in PCs. Then to develop video applications and video analysis software, the software platform of the system is also introduced for rapid prototyping of video applications and video analysis software. By using the constructed testbed, we are now tackling multimedia indexing and mining techniques towards inter-video structuring in huge multimedia corpora.

References

1. Yukinobu Taniguchi, Akihito Akutsu, Yoshinobu Tonomura, and Hiroshi Hamada, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing," in *Proc. of ACM Multimedia*, 1995, pp. 25–33.
2. Yuichi Nakamura and Takeo Kanade, "Semantic analysis for video contents extraction — spotting by association in news video," in *Proc. of ACM Multimedia 97*, 1997.
3. Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE MultiMedia*, vol. 6, no. 1, pp. 22–35, January-March (Spring) 1999.
4. Michael G. Christel, "Visual digests for news video libraries," in *Proc. of ACM Multimedia*, 1999, pp. 303–311.
5. Xinbo Gao and Xiaou Tang, "Unsupervised and model-free news video segmentation," in *Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries*, 2001, pp. 58–64.
6. N. Babaguchi, S. Sasamori, T. Kitahashi, and R. Jain, "Detecting events from continuous media by intermodal collaboration and knowledge use," in *Proc. of International Conference on Multimedia Computing and Systems (ICMCS)*, 1999, pp. 782–786.
7. Reiko Hamada, Shin'ichi Satoh, Shuichi Sakai, and Hidehiko Tanaka, "detection of important segments in cooking videos," in *Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries*, 2001, pp. 118–123.
8. Howard D. Wactlar, Michael G. Christel, Yihong Gong, and Alexander G. Hauptmann, "Lessons learned from building a terabyte digital video library," *IEEE Computer*, vol. 32, no. 2, pp. 66–73, 1999.
9. I. Ide, H. Mo, N. Katayama, S. Satoh, "Topic threading for structuring a large-scale news video archive", *Int. Conf. on Image and Video Retrieval (CIVR2004)*, LNCS vol.3115, 2004, pp.123–131.
10. H. Mo, F. Yamagishi, I. Ide, N. Katayama, S. Satoh, and M. Sakauchi, "Key Image Extraction from a News Video Archive for Visualizing its Semantic Structure," *PCM 2004* (to appear).
11. C. Yu, H. Mo, N. Katayama, S. Satoh, and S. Asano, "Semantic Retrieval in a Large-Scale Video Database by Using both Image and Text Feature," *PCM 2004* (to appear).
12. F. Yamagishi, S. Satoh, and M. Sakauchi, "A News Video Browser Using Identical Video Segment Detection," *PCM 2004* (to appear).
13. S. Satoh and N. Katayama, "An Efficient Implementation and Evaluation of Robust Face Sequence Matching," *Proc. of ICIA99*, 1999, pp. 266–271.

Sample Selection Strategies for Relevance Feedback in Region-Based Image Retrieval

Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa

INRIA Rocquencourt, 78153 Le Chesnay Cedex, France
{Marin.Ferecatu, Michel.Crucianu, Nozha.Boujemaa}@inria.fr

Abstract. The success of the relevance feedback search paradigm in image retrieval is influenced by the selection strategy employed by the system to choose the images presented to the user for providing feedback. Indeed, this strategy has a strong effect on the transfer of information between the user and the system. Using SVMs, we put forward a new active learning selection strategy that minimizes redundancy between the examples. We focus on region-based image retrieval and we expect our approach to produce better results than existing selection strategies. Experimental evidence in the context of generalist image databases confirms the effectiveness of this selection strategy.

1 Introduction

The concept of *semantic gap* has been extensively used in the Content Based Image Retrieval (CBIR) research community to express the discrepancy between the low-level features that can be readily extracted from the images and the descriptions that are meaningful to the users of the search engines [1].

Image regions are usually perceived as being closer to semantic concepts and one way to address the semantic gap is to concentrate the search at the region level where a relation between concepts and regions is easier to establish [2]. Another solution for reducing the semantic gap is to cut a search session into several consecutive retrieval rounds (iterations) and let the user provide feedback regarding the results of every retrieval round, e.g. by qualifying images returned as either “relevant” or “irrelevant” [3] (relevance feedback, RF).

In order to maximize the ratio between the quality (or relevance) of the results and the amount of interaction between the user and the system, the selection of images for which the user is asked to provide feedback at the next round must be carefully studied. For a *target search* scenario, where the user is searching for a specific image, interesting ideas were introduced in [4]. At every round, the user is required to choose between two images presented by the engine and the selection strategy must let the user remove a maximal amount of uncertainty regarding the target. We consider that this criterion translates into two complementary conditions for the images in the selection: (1) each image must be ambiguous given the current estimation of the target and (2) the redundancy between the different images has to be low. However, computational

optimizations are required for searching larger sets of images (*category search*) and for selecting more than 2 images.

Based on the definition of active learning (see for example [5]), the selection of examples for training SVMs to perform general classification tasks is studied in [6]. In the early stages of learning, the classification of new examples is likely to be wrong, so the fastest reduction in generalization error can be achieved by selecting the example that is farthest from the current estimation of the frontier. During late stages of learning, the classification of new examples is likely to be right but the margin may be suboptimal, so the fastest reduction in error can be achieved by selecting the example that is closest to the current estimation of the frontier. Note that, according to the classical formulation of active learning, the authors only consider the selection of single examples for labeling (or addition to the training set) at every round.

Also for SVM learners, several selection criteria are presented in [7] and applied to content-based text retrieval with relevance feedback. The simplest (and computationally cheapest) of these criteria consists in selecting the texts whose representations (in the feature space induced by the kernel) are closest to the hyperplane currently defined by the SVM. We shall call this simple criterion the selection of the “most ambiguous” (MA) candidate(s). This selection criterion is justified in [7] by the fact that knowledge of the label of such a candidate halves the version-space (the set of learner parameters that are compatible with the already labeled examples). In order to minimize the number of feedback rounds, the user is asked to label several examples at every round and all these examples are selected according to the MA criterion. In [8] the MA selection criterion is applied to CBIR with relevance feedback and shown to produce a faster identification of the target images than the selection of random images for further labeling.

In the next section we put forward a new active selection strategy based on the reduction of the redundancy between the examples presented to the user. Experimental evidence in Section 3 shows that our strategy performs well compared with other strategies in generalist database region based query contexts. Concluding remarks are given in Section 4.

2 Reduction of the Redundancy

While the MA criterion provides a computationally effective solution to the selection of the most ambiguous images (satisfying the first condition mentioned above), when used for the selection of more than one candidate image it does not remove the redundancies between candidates.

We suggest here to translate this condition of low redundancy into the following additional condition: if x_i and x_j are the input space representations of two candidate images, then we require a low value for $K(x_i, x_j)$ (i.e. of the value taken by the kernel for this pair of images). If the kernel K is inducing a Hilbert structure on the feature space, if $\phi(x_i)$, $\phi(x_j)$ are the images of x_i , x_j in this feature space and if all the images of vectors in the input space have

constant norm, then this additional condition corresponds to a requirement of quasi-orthogonality between $\phi(x_i)$ and $\phi(x_j)$ (since $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$). We shall call this criterion the selection of the “most ambiguous and orthogonal” (MAO) candidates.

We note that the MAO criterion can be extended to reduce redundancies between the examples selected during subsequent RF rounds. This additional constraint may be important in situations where the number of labeled examples is much lower than the dimension of the input space and the classes are restricted in most directions.

The MAO criterion has a simple intuitive explanation for kernels $K(x_i, x_j)$ that decrease with an increase of the distance $d(x_i, x_j)$ (which is the case for most common kernels): it encourages the selection of unlabeled examples that are far from each other in input space, allowing to better explore the current frontier.

To implement this criterion, we first perform an MA selection of a larger set of unlabeled examples. Then, we build the MAO selection by iteratively choosing as a new example the vector x_j that minimizes the highest of the values taken by $K(x_i, x_j)$ for all the x_i examples already included in the current MAO selection. This can be written as:

$$x_j = \operatorname{argmin}_{x \in S} \max_i K(x, x_i)$$

where S is the set of images not yet included in the current MAO selection and $x_i, i = 1 \dots n$ are the already chosen candidates.

In a general classification context, a similar “diversity” condition for the selected examples was put forward in [9] and evaluated on several benchmark classification problems from the UCI database. The condition is justified by reference to the version space account suggested in [7]: diversity is maximized when the hyperplanes associated to the individual examples are orthogonal and are thus complementary to each other in halving the version space.

We note that the MA criterion in [7], [8] is the same as the one put forward in [6] for the late stages of learning. This clarifies the fact that the MA criterion relies on two important further assumptions: first, the prior on the version space is rather uniform; second, the solution found by the SVM is close to the center of gravity of the version space.

In early stages of the learning the frontier is very unreliable and selecting those unlabeled examples that are currently considered by the learner as (potentially) the most relevant can sometimes produce a faster convergence of the frontier during the first few rounds of RF.

For this reason, we added to our comparisons the following criteria: select the “most positive” unlabeled examples according to the current decision function of the SVM (denoted as MP criterion) and select the “most positive and orthogonal” unlabeled examples (denoted as MPO). The MPO criterion adds to MP the condition of low redundancy previously described. When comparing the MP criterion to the suggestion in [6] for the early stages of learning, we see that we only focus on the examples for which the values taken by the decision function

of the SVM are maximal and completely ignore the examples for which these values are minimal; this is because of the asymmetry of the retrieval context: in general, the number of relevant items is expected to be much lower than the number of irrelevant items.

3 Experimental Evaluation

To test our selection criterion we selected a groundtruth database of image regions, built from a generalist image database. The first stage was to automatically obtain a coarse segmentation of the images in the database using the algorithms described in [2].

To describe the visual content of image regions we used Laplacian wighted histograms, probability weighted histograms, shape histograms based on the Hough transform and classic HSV color histograms. Weighted color histograms rely on the idea that not all pixels are equal when it comes to their contribution to the histogram. Pixels from uniform regions of an image are less important than pixels from regions where there are important changes in color (see [10]). These histograms integrate a local measure of the uniformity of the pixels, and thus have a texture description value added to their primary intended color description.

The final feature vector is the concatenation of individual feature vectors and has more than 600 dimensions. The very high number of dimensions of the joint feature vector can make RF impractical even for medium-size databases. Also, the higher the dimensionality of the description space, the more difficult is the task of the learner. We use a linear PCA to reduce the dimension of the feature vector more than 5 times, without a significant loss (less than 5%) on the precision/recall diagrams in a query by example context.

To build the groundtruth, we annotated by hand 5401 regions from a total of 44286 automatically segmented regions in our database. We obtained 27 classes (such as hair, face, sea, forrest, village, shutters, etc.), most of them containing more than 150 regions. The results we show here correspond to the “sea” class (510 image regions) and to the “face” class (523 image regions).

At every feedback round the (emulated) user must label as “relevant” or “irrelevant” all the images in a window of size $ws = 9$. A search session is initialized by considering one “relevant” example and $ws - 1$ “irrelevant” examples. Every image in every class serves as the initial “relevant” example for a different RF session, while the associated initial $ws - 1$ “irrelevant” examples are randomly selected.

For the SVM we employed the triangular kernel, $K(x_i, x_j) = -\|x_i - x_j\|$, because in all our experiments it performed better than other kernels (RBF, Laplace). Also, this kernel has the property of making the frontier found by SVMs invariant to the scale of the data (see [11]). Classes of image regions in generalist databases usually have very different scales in the space of low-level descriptors; thus, kernels producing scale invariance are to be preferred to the classical ones that are very sensitive to a scale parameter.

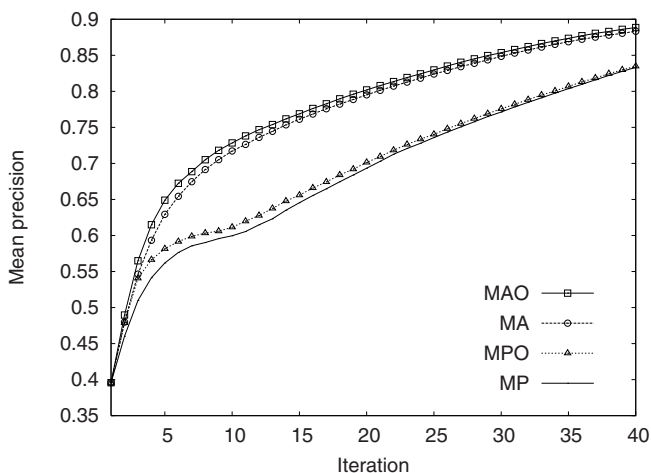


Fig. 1. Comparison of the selection strategies for the “sea” class.

We first evaluated the different selection criteria in a **ranking scenario**: finding items in a specific target set, by focusing on ranking most of the “relevant” images before the “irrelevant” ones rather than on finding a frontier between the class of interest and the other images. In order to evaluate the speed of improvement of this ranking, we must use a measure that does not give a prior advantage to one selection criterion. For example, by taking into account already labeled images plus those selected for being labeled during the current round, we should obviously favor the MP and MPO criteria over MA and MAO. We decided to use instead the following precision measure: at every RF round, we count the number of “relevant” images found in the first N images considered as most positive by the current decision function of the SVM (N being the number of images in each class).

The evolution of the mean precision during successive RF iterations (rounds) is presented in Fig. 1 for the “sea” class and in Fig. 2 for the “face” class. The “mean precision” value shown is obtained as the mean over all the image regions in the class of the precision measure described above. Clearly, the reduction of the redundancy between the images selected for labeling improves the results, both for MAO with respect to MA and for MPO with respect to MP. Also, in these cases the MA and MAO selection criteria compare favorably to the MP and MPO criteria.

The **second type of scenario** we evaluated consists in finding a frontier between “relevant” and “irrelevant” images, which can be important for extending textual annotations of some images in the “relevant” class to the others. In this case, we have to evaluate the speed of improvement of the classification. The classification error is defined here as $n/N + (N - p)/N$, where N is the class size, n is the number of false positives and p is the number of true positives (thus $N - p$ is the number of false negatives). In Fig. 3 we can see the evolution of

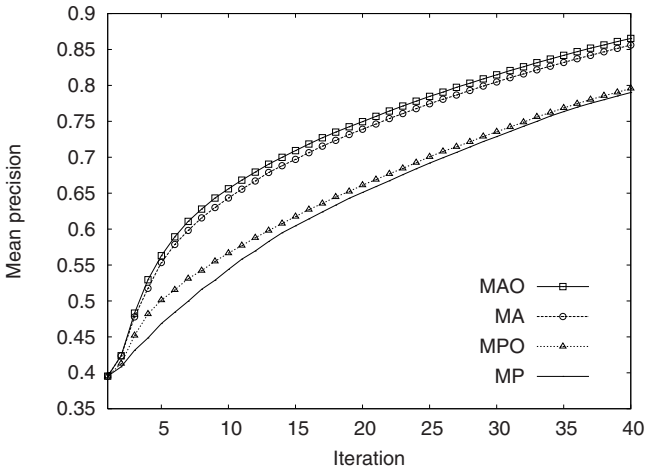


Fig. 2. Comparison of the selection strategies for the “face” class.

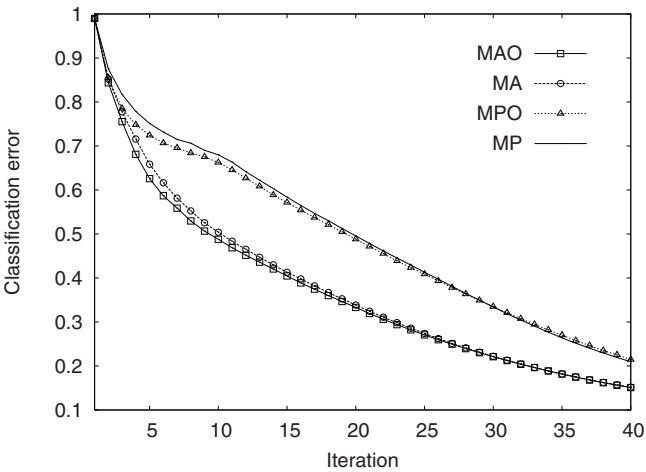


Fig. 3. Evolution of the classification error obtained with different selection strategies for the “sea” class.

the classification error obtained with the different selection criteria for the “sea” class. As expected, the convergence is fastest for the MAO selection criterion, followed by the MA criterion.

Similar results are obtained for most of the other classes, with one remark: simple classes, with a small number of elements concentrated in a relatively small region in feature space, have an almost identical behavior for all the selection strategies. Also, we found only two classes (out of the 27) where the MP criteria performed slightly better than the MA ones: “building” and “city”.



Fig. 4. Searching for village regions in a generalist database. The contours of the regions matching the query are in red (dark gray) and those of the other regions in light blue (light gray).

Principal component analysis reveals that the projections of the two classes are rather spread and biased discriminant analysis shows that these classes are very mixed, making them very difficult to separate by the SVM. This suggests that complementary image features should be used to discriminate the two classes.

4 Conclusion

In content-based image retrieval with relevance feedback, the strategy employed by the search engine for selecting the images presented to the user at every feedback round is very important for the transfer of information between the user and the system. Using SVMs as learners, we put forward an improved active learning selection strategy, based on a reduction of the redundancy between the images selected at every feedback round.

By comparing this strategy to alternative strategies for the retrieval of regions in the context of generalist image databases, we have shown that it performs better in ranking most of the “relevant” images before the others and also speeds up the convergence of the frontier around the class of interest. This last aspect is important when relevance feedback is used as a tool for extending semantic annotations.

As a visual example of retrieval, in Fig 4 we present the third screen of results returned by our CBIR system IKONA (see [10]), on the database used in this paper. In this example the class the user is searching for is “village” (having 87

examples) and at this point of the RF session there are 4 positive examples and 9 negative examples already annotated.

In this case, all the regions returned belong to the “village” class; nevertheless they have different characteristics: some are close-ups and some are not, the texture is very different, some focus on single buildings and some are global views, etc. Moreover, this class is easy to confuse with other classes (“rock”, “mountain”, “city”) in terms of color, texture and shapes, characteristics used by the image features.

References

1. Gevers, T., Smeulders, A.W.M.: Content-based image retrieval: An overview. In Medioni, G., Kang, S.B., eds.: *Emerging Topics in Computer Vision*, Prentice Hall (2004)
2. Fauqueur, J., Boujemaa, N.: Region-based image retrieval: Fast coarse segmentation and fine color description. *Journal of Visual Languages and Computing (JVLC)*, special issue on Visual Information Systems **15** (2004) 69–95
3. Zhou, X.S., Huang, T.S.: Relevance feedback for image retrieval: a comprehensive review. *Multimedia Systems* **8** (2003) 536–544
4. Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T., Yianilos, P.N.: The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing* **9** (2000) 20–37
5. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. *Journal of Artificial Intelligence Research* **4** (1996) 129–145
6. Campbell, C., Cristianini, N., Smola, A.: Query learning with large margin classifiers. In: *Proceedings of ICML-00, 17th International Conference on Machine Learning*, Morgan Kaufmann (2000) 111–118
7. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: *Proceedings of ICML-00, 17th International Conference on Machine Learning*, Morgan Kaufmann (2000) 999–1006
8. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: *Proceedings of the 9th ACM international conference on Multimedia*, ACM Press (2001) 107–118
9. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: *Proceedings of ICML-04, International Conference on Machine Learning*. (2003) 59–66
10. Boujemaa, N., Fauqueur, J., Ferecatu, M., Fleuret, F., Gouet, V., Saux, B.L., Sahbi, H.: Ikona: Interactive generic and specific image retrieval. In: *Proceedings of the International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR'2001)*. (2001)
11. Fleuret, F., Sahbi, H.: Scale-invariance of support vector machines based on the triangular kernel. In: *3rd International Workshop on Statistical and Computational Theories of Vision*. (2003)

Comparison of Two Different Approaches to Detect Perceptual Noise for MPEG-4 AAC

Cheng-Hsun Yu¹ and Shingchern D. You²

¹ Network Video Business Unit,
AVerMedia Technologies Inc.,
7-10F, 135, Jian-Yi Rd.,
Chung-Ho city, Taipei Hsien, Taiwan

² Department of Computer Science and Information Engineering,
National Taipei University of Technology,
1, Sec. 3, Chung-Hsiao East Rd.,
Taipei 106, Taiwan
`you@csie.ntut.edu.tw`

Abstract. The MPEG-4 AAC introduces a new tool called Perceptual Noise Substitution (PNS) whose function is to use locally generated noise to replace the noise inside the coded music. The key issue to the PNS tool relies on the correctly detect the perceptual noise. In this paper, we point out the potential problem of the approach suggested in the ISO's document and compare it with the other approach, proposed by the original contributors of the PNS concept, and discuss their relative performance.

1 Introduction

Perceptual audio coding is the mainstream of audio coding now. Audio coding standards such as MPEG-1 [1], MPEG-2 [2],[3], and AC-3 [4], all are in this category. In the MPEG-2 standard, two audio coding schemes are available, namely part 3 [2] and part 7 [3]. The part 3 is designed to be MPEG-1 back compatible (BC); whereas the part 7 is not. That is the reason that the part 7 was originally known as MPEG-2 NBC standing for Non-Back Compatible. The part 7 was finally named as MPEG-2 Advanced Audio Coding (AAC). Subjective (listening) experiments showed that the coding quality of AAC was better than that of MPEG-2 BC [5]. Therefore, the MPEG-2 AAC has gained more attention.

With the increasing demand at low bitrate coding, the MPEG committee later developed the MPEG-4 natural audio coding [6], which was largely based on the MPEG-2 AAC scheme with some additional powerful new tools. Among the new tools, the Perceptual Noise Substitution (PNS) [8,9] tool has been shown to be effective for coding signals at low bitrates. The idea of this tool is simple: all noise sounds alike. Therefore, if a piece of music is perceptually similar to noise, we may, instead of coding the waveform of the signal, just encode and

transmit the information of average power of the signal. Thus, the saved bits may be allocated to other signal instances demanding more bits to encode.

Although the concept of PNS is simple, it is not easy to find a segment (or block) of noise-like signal. Therefore, it would be more practical to replace by noise only a certain scalefactor bands, instead of the whole signal spectrum. That is the basic operating concept of the MPEG-4 PNS tool. In addition, low frequency components of the signal should not be substituted, or obvious artifact will be noticed. Therefore, only high frequency components are subject to be substituted.

With only high frequency bands are subject to be substituted, the number of bits saved using PNS actually is very few. Based on our experiments, we find that the improvement of subjective quality with PNS is not due to re-allocation of saved bits. Rather, it is due to adding noise to high frequency bands. Usually a signal coded at 32 kbps sounds very rough because of not having enough high frequency components. With PNS, the band-limited noise is added to the decoded signal. This noise provides the feeling that the sound contains abundant high-frequency components. As shown in Fig. 1, a piece of signal with rich high frequency components lost most of its high frequency components when coded at 32 kbps. With PNS, randomly generated high frequency components are added back to the coded signal. Thus, the signal becomes perceptually better.

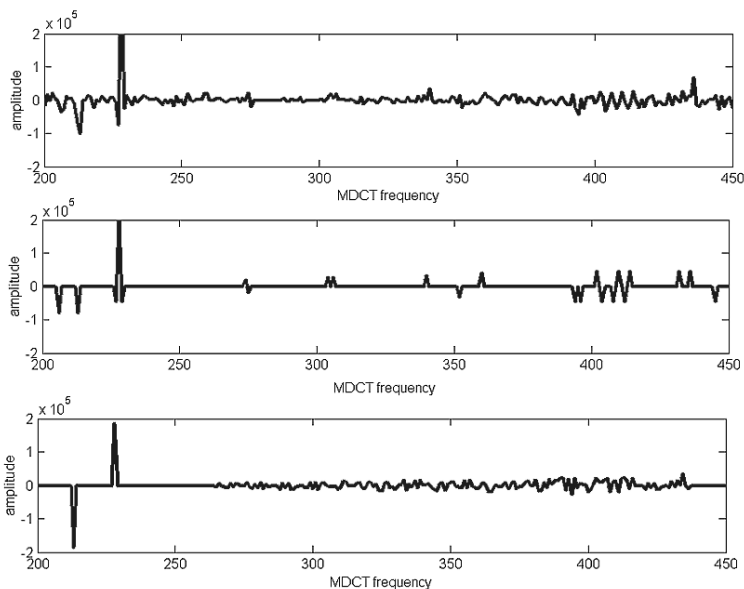


Fig. 1. The MDCT spectrum of a piece of music. Top: the spectrum of the original signal; middle: the spectrum of the signal coded at 32 kbps; bottom: the spectrum of the coded signal with PNS activated.

2 Detection of Perceptual Noise

For the PNS tool to be effective, it relies on the accurate detection of noise-like spectrum in high-frequency scalefactor bands. The quality of the coded signal will be even worse than not using PNS if falsely substituting noise for a piece of a music with strong harmonic components. Thus, it deserves to pay more attention to the noise detection strategy.

The original contributor of PNS therefore proposed several approaches [9] to detect perceptual noise. Among the approaches, the one named “noise detection by prediction in subbands” is more suitable for implementation in MPEG-4 AAC. In the following, we refer to this one as the Schulz’s approach with a brief description in the following. Incoming PCM samples pass through an analysis filterbank to split the signal into various sub-bands. A linear predictor is then used to predict samples in each sub-band. The predicted results are then combined together through a synthesis filterbank. Since it is impossible to predict pure noise, a large prediction error in a sub-band indicates that the sub-band is noise-like. In Schulz’s method, the original and the synthesized signals are then individually taken FFT’s. The tonality is determined by the normalized differences between the original and the synthesized spectral lines. In our case, we use MDCT in place of FFT to fit into the MPEG-4 architecture. Therefore, the MDCT of the predicted signal is computed. The MDCT of a delayed version of the original signal is also computed. A scalefactor band is noise-like if the difference between the predicted one and the original one is large in that band. Although Schulz gave the generic method to detect perceptual noise, the paper did not provide implementation details such as the number of subbands used in the experiment. To be able to judge the performance of this approach, we use a PQF (Polyphase Quadrature Filterbank) as the analysis and filterbank in our implementation, as shown in Fig. 2, to be used later in the experiments.

Despite the availability of the Schulz’s approach, the MPEG-4 standard (Annex 4.B.12) suggests another method to determine whether a scalefactor band is noise-like or not. A scalefactor band is assumed to be noise-like if the following two conditions are satisfied: (i) the tonality index calculated in the psychoacoustic model is smaller than a threshold; (ii) the energy in the band does not greatly change over time. Since the tonality index is a by-product of psychoacoustic model and the calculation of the energy in a band is simple, this approach is simple and requires much low computational burden compared to the Schulz’s approach.

Although the ISO’s approach is simple, it is by no means to be better. The tonality index in ISO’s approach is calculated based on whether a spectral value of the FFT can be linearly predicted by two previous values in the same spectral line from two preceding blocks. However, if the incoming signal has a small variation in frequency, the prediction in spectral lines will fail. Therefore, this value is very sensitive to frequency variation.

Since there are two different approaches available, it deserves to judge the relative performance between these two. Therefore, we conducted some experiments for this purpose.

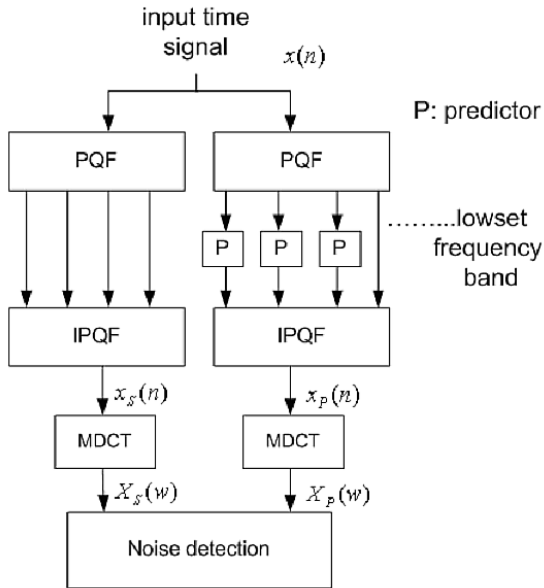


Fig. 2. Detection of perceptual noise using PQF and linear predictors.

To implement the Schulz's approach, we have tried to predict the incoming signal directly without using filterbank. The results were poor. This confirms that an analysis filterbank must be used to limit the bandwidth of the predicted samples. Therefore, we used a four-band PQF as the filterbank as mentioned previously. The block diagram is shown in Fig. 2. The PQF used here is the same one as used in the scalable profile in the MPEG-2 AAC. The reason for using this filterbank is because of its simplicity and relatively lower complexity. Besides, its code (program) is available in the reference software [7]. The predictor used in each sub-band is a 30th-order linear predictor with its coefficients obtained using the Levinson-Durbin algorithm. In contrast to the Schulz's approach, the ISO's approach is available in the reference software. Therefore, we use it as the comparison counterpart.

3 Experiments and Results

Several experiments have been conducted to examine the relative performance of these two approaches. These experiments belonged to two different parts. The first part determined whether the above two approaches could distinguish two different types of signals. The second part was a subjective (listening) experiment.

The first experiment in the first part was to check if the tonality index could distinguish a 6-kHz sine wave from a narrow-band FM signal (from 6 kHz to 6.1

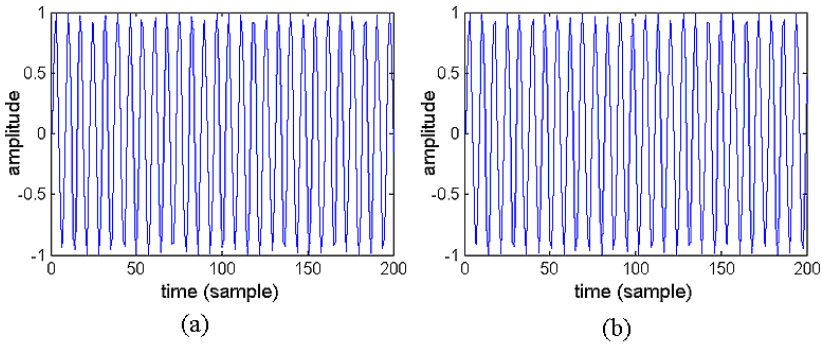


Fig. 3. (a) A 6 kHz sine wave; (b) A narrow-band FM from 6 kHz to 6.1 kHz.

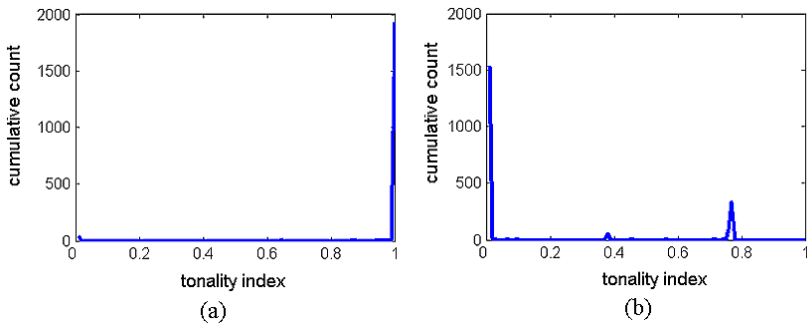


Fig. 4. (a) The histogram of tonality index with a 6-kHz sine wave as the input using the ISO's approach; (b) the histogram with a slow-varying signal.

kHz), as shown in Fig. 3. The results are depicted in Fig. 4. In the figure, the horizontal axis indicates the value of tonality index and the vertical axis is the number of scalefactors, whose tonality index are given in the horizontal axis. This types of plots is also known as histograms. Fig. 4 shows that if a pure sine wave is coded then the tonality indices associated with most scalefactor bands are almost one, which means that the incoming signal has strong tone components. However, for a narrow-band FM signal, despite its similarity to a pure tone, the tonality indices of most scalefactor bands approach zero, indicating that the incoming signal is noise-like. This confirms our previous claim: the tonality index, due to using prediction in the frequency domain, is very sensitive to the frequency variation. The same signals are used to test the discrimination power of the Schulz's approach, and the results are given in Fig. 5. In the figure, the normalized prediction error is used in place of tonality index. This error is the difference between two sets MDCT outputs in a scalefactor band. Later, we shall use this value to determine if this band is noise-like. Conceptually, a pure tone

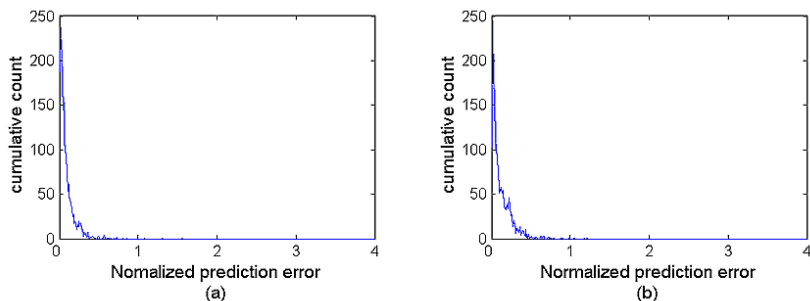


Fig. 5. (a) The histogram of normalized prediction error with a 6-kHz sine wave as the input using the Schulz's approach; (b) the histogram with a slow-varying signal.

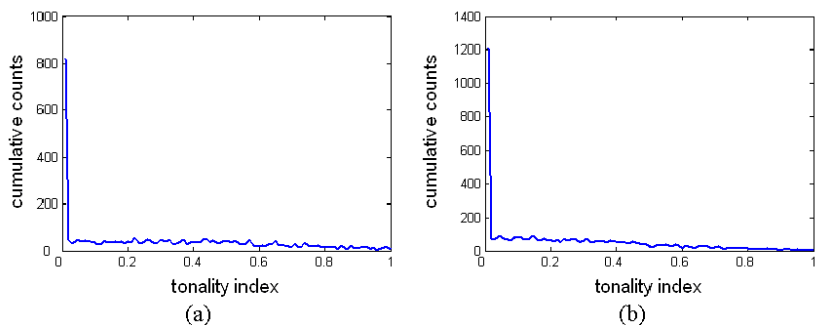


Fig. 6. The histograms of the ISO's approach with different types of music. (a) Female vocal; (b) Orchestra.

has a low prediction error. Note that a pure tone possesses a high tonality index in the ISO's method. From Fig. 5 it is observed that this approach is not sensitive to a small frequency variation.

The second experiment was similar to the first one but this time two different types of music were used. The first one was a sound track of female vocal, and the second one orchestra music. The histograms of tonality index of both types of music are shown in Fig. 6 using the ISO's approach. It is hard to tell the difference by viewing the histograms. The results based on the Schulz's approach are shown in Fig. 7. Apparently, the two histograms are different. Based on the first and second experiments, we conclude that the discrimination power of the tonality index is very limited. Therefore, in the ISO's approach, whether PNS is used is solely based on the energy difference between scalefactor bands in consecutive blocks.

The subjective experiment was also carried out. Since PNS is primarily used at low bitrates, the signals were coded at 32 kbps in the experiment. With this

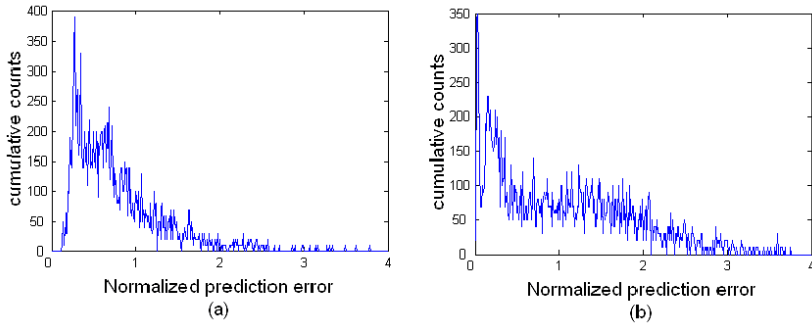


Fig. 7. The histograms of the Schulz’s approach with different types of music. (a) Female vocal; (b) Orchestra.

Table 1. Types of music used in the experiments.

song	Female vocal	music	Soft music
orchestra	Waltz orchestra	samprg	Rock and roll
else	Soft music	castanets	Castanets

Table 2. Experimental results for two different approaches.

Music name	Schulz’s approach better	Both equal	ISO’s approach better	Average score
song	6	8	1	0.33
music	1	11	3	-0.13
orchestra	0	14	1	-0.06
samprg	5	8	2	0.2
else	3	10	2	0.06
castanets	10	4	1	0.6

low bitrate, the PEAQ [10] program was not used because it was mainly used for rating high-quality coded signals. Due to lacking of experienced audiences, we used a simplified CMOS (Comparative Mean Opinion Score) method in the experiment. Fifteen grad students were asked to give opinions after listening to three pieces of music arranged in Ref/A/B format, where Ref was the original signal, and A and B were the two coded results. The opinions were: (i) A is better than B; (ii) A is equal to B; (iii) A is worse than B. The signal coded using either approach was randomly assigned to either A or B. Besides, the audiences had no knowledge about which one was coded by which method. The signals for comparison were six pieces of different music. The contents of the signals are listed in Table 1. Because conducting subjective experiments require a lot of time and man power, it is a common practice to use only a small amount of songs in the subjective test. For example, [9] used only twelve different songs.

The results of the subjective test, as given in Table 2, indicate that the Schulz's approach is slightly better on the average than its counterpart. In addition, it performs much better for the music of castanets. Because castanets has fast energy changes, the ISO's method has difficulties to detect the noise-like components of the music. Overall speaking, the ISO's approach is simple but not good enough. Thus, how to modify the ISO's approach for better performance requires further study.

4 Conclusions

In this paper, we have compared two different approaches to detect perceptual noise for the PNS tool provided in the MPEG-4 AAC. The first approach, given by Schulz, uses predictors in the time domain, and the second one, given in the ISO's standard as an annex, uses the tonality index obtained from the psychoacoustic model as the decision basis. The experimental results show that the Schulz's approach is better for some types of music with the price of much higher computational complexity than the other one.

References

1. ISO/IEC: Information technology - coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 3: Audio. IS 11172-3 (1993)
2. ISO/IEC: Information technology - Generic coding of moving pictures and associated audio information - Part 3: Audio. 2nd Ed. IS 13818-3 (1998)
3. ISO/IEC: Information technology - generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC). IS 13818-7 (1997)
4. Advanced Television Systems Committee: Digital audio compression standard (AC-3). Doc. A/52, (1995)
5. Bosi M., et al: ISO/IEC MPEG-2 advanced audio coding. *Journal of Audio Eng. Soc.* 45 (1997) 789 - 812
6. ISO/IEC: Information technology - Coding of audio-visual objects - Part 3: Audio, subpart 4 general audio coding. IS 14496-3 (1999)
7. ISO/IEC: Information technology - Coding of audio-visual objects - Part 5: Reference software for MPEG-4. IS 14496-5 (2001)
8. Herre J., D. Schulz: Extending the MPEG-4 AAC coding by perceptual noise substitution. 104st Conference of the Audio Eng. Soc., Amsterdam, CA (1998) pre-print 4720
9. D. Schulz: Improving audio codecs by noise substitution. *Journal of Audio Eng. Soc.* 44 (1996) 593 - 598
10. T. Thiede, et al: PEAQ—The ITU standard for objective measurement of perceived audio quality. *Journal of Audio Eng. Soc.* 48 (2000) 3 - 29

Optimum End-to-End Distortion Estimation for Error Resilient Video Coding

Yuan Zhang^{1,2}, Qingming Huang¹, Yan Lu³, and Wen Gao¹

¹ Graduate School, Chinese Academy of Sciences, 100080 Beijing, China
{yzhang, qmhuang, wgao}@jdl.ac.cn

² Beijing Broadcasting Institute, 100024 Beijing, China

³ Microsoft Research Asia, 100080 Beijing, China
t-yanlu@microsoft.com

Abstract. End-to-end distortion estimation plays an important role in error-resilient video coding. The intuitive method is to simulate the decoding process many times at the encoder, as used in the H.264 test model. However, the computing complexity is very high. In this paper, an optimum end-to-end distortion estimation scheme is proposed. Concretely, the correlations of the potentially propagated errors of the different frames are modeled based on the theoretical analysis, and then a block-based potential distortion tracking scheme with very low computing complexity is proposed. Statistics show that the gaps of the estimated distortions between the proposed algorithm and the H.264 test model become smaller and smaller when the times of simulated decoding in H.264 test model increases. In other words, the proposed algorithm is more accurate than H.264 test model. Moreover, an improved rate-distortion optimization algorithm based on the optimum end-to-end distortion estimation is proposed, wherein the rate control is also jointly utilized.

1 Introduction

Transmitting the hybrid-coded video over the packet-switched networks with packet losses often suffers from the error propagation and leads to the well-known drifting phenomenon [1]. To tackle this problem, error-resilient video coding and error concealment algorithms have been devised at the encoder and the decoder, respectively [2]. Intra block refreshment is one of the most popular error-resilient coding schemes. In standard-compliant techniques, intra coding can suppress the error propagation at the cost of reduced coding efficiency. The main problem is about how to achieve the optimum transmission efficiency while considering both the absolute coding efficiency and the suppression of potential errors. Towards this goal, many researchers have been focused on the methods of adaptively inserting intra blocks into the coded frame/sequence.

The early algorithms were developed to randomly place intra MBs [3], or periodic intra-code contiguous blocks in a frame [4]. The intra refresh frequency was determined in a heuristic way and the intra-coding was applied uniformly to the whole frame. Then the content-adaptive coding mode selection scheme

was proposed to intra-code the MBs at regions with high activity [5]. To further improve the performance by jointly considering the network condition and the error concealment, rate-distortion (RD) optimized algorithms have been proposed with the theme of achieving an optimum trade-off between the distortion and the bit rate. Rate distortion optimization (RDO) scheme is well known in source coding, in which the distortion only refers to the quantization errors. However, in the error-prone environment, transmission errors inevitably increase the actual distortion of the reconstructed frame. Therefore, the channel distortion should also be considered in the RDO-based video coding.

The recent work proposed in [6] developed a statistical model to estimate the channel distortion and decide the intra refresh rate before coding each frame. However, the sequence/frame-level adaptive intra-refreshing algorithm considers only the overall R-D behavior of the whole video/frame, which lacks the accuracy in the local end-to-end distortion estimation. In [7], a recursive optimal per-pixel estimate (ROPE) algorithm is proposed to estimate the end-to-end distortion at pixel level. Although it works well for the H.263 codec, its extension to the up-to-date H.264 recommendation is not straightforward due to the high complex prediction schemes employed in H.264, such as intra-prediction, in-loop filter and sub-pixel motion compensation. Consequently, an error robust rate distortion optimization (ER-RDO) method has been developed in the H.264 test model for video coding in packet loss environment [8][9]. The decoded MB distortion is computed as the average over the K distortions by decoding this MB K times based on the erroneous reference frames. The expected decoder distortion can be estimated accurately in the encoder if K is chosen large enough. However, the high computational complexity and implementation cost make it impractical when K is increased.

In this paper, we propose an optimum end-to-end distortion estimation scheme for error-resilient video coding. The basic end-to-end distortion model is derived according to the theoretical analysis at the pixel level. The correlations of the potential distortions among different frames are derived. Then, the block-level implementation based on the theoretical model is proposed to tackle the influence of sub-pixel motion compensation. In the proposed scheme, a distortion map is defined to store the potential errors of each frame that may propagate to its future frames. In other words, when coding the current frame, the potential channel distortions of its reference frames have been known as a priori. Based on the proposed end-to-end distortion estimation scheme, an improved RDO-based intra/inter mode selection algorithm is developed, in which a new Lagrange parameter in RDO coding is employed. Since the video codec is usually associated with some rate control scheme in application, the rate control technique in H.264 test model is jointly implemented with the proposed error resilient video coding scheme.

The rest of this paper is organized as follows. The end-to-end distortion model is described in Section 2. Simulation results as well as the discussion are also presented. In Section 3, a joint rate-distortion optimization method for H.264

video encoding in packet loss environment is described. Afterwards, experimental results are presented. Section 4 concludes this paper.

2 End-to-End Distortion Estimation

2.1 Theoretical Model

Let f_n^i be the original value of pixel i in the n th video frame, and \widehat{f}_n^i be the corresponding reconstruction value at the encoder side. Suppose \widehat{f}_{ref}^j and \widehat{r}_n^i denote the predicted value (i.e. pixel j in reference frame ref) and the quantized residual, respectively. Then \widehat{f}_n^i is given by $\widehat{f}_n^i = \widehat{f}_{ref}^j + \widehat{r}_n^i$. Let \widetilde{f}_n^i be the reconstructed value at the decoder, which is a random variable for the encoder. Assume that the temporal error concealment is used to reconstruct this pixel in case of packet loss. Let this replacement be the pixel k in the previous frame $n - 1$, denoted as \widetilde{f}_{n-1}^k . If packet loss rate is p , then we have:

$$\widetilde{f}_n^i = (1 - p)(\widetilde{f}_{ref}^j + \widehat{r}_n^i) + p\widetilde{f}_{n-1}^k. \quad (1)$$

Therefore, the expected overall distortion of pixel i in frame n at the decoder is:

$$\begin{aligned} d_n^i &= E\left\{(f_n^i - \widetilde{f}_n^i)^2\right\} = (1 - p)E\left\{(f_n^i - \widehat{r}_n^i - \widetilde{f}_{ref}^j)^2\right\} + pE\left\{(f_n^i - \widetilde{f}_{n-1}^k)^2\right\} \\ &\approx (1 - p)E\left\{(f_n^i - \widehat{f}_n^i)^2\right\} + (1 - p)E\left\{(\widehat{f}_{ref}^j - \widetilde{f}_{ref}^j)^2\right\} + pE\left\{(f_n^i - \widetilde{f}_{n-1}^k)^2\right\} \\ &= (1 - p)d_s + (1 - p)d_{ep_ref} + pd_{ec}, \end{aligned} \quad (2)$$

where d_s denotes the quantization distortion, d_{ep_ref} denotes the error propagated distortion from the predicted pixel in the reference frame, and d_{ec} denotes the error concealment distortion for pixel i in frame n . For a pixel in the intra-coded MB, its predict range is restricted in the current slice. If the MB is not lost, its predictor is not lost too. Therefore, the reconstructed pixel value at the encoder and the decoder are the same and therefore d_{ep_ref} is zero. For a pixel in the inter-coded MB, even in case that there are not any channel errors, it still suffers from the propagated distortion through the motion compensation path. Now the key point is about how to obtain the error-propagated distortion. Obviously, after the current frame has been encoded, the error-propagated distortion d_{ep} can be achieved by:

$$\begin{aligned} d_{ep} &= E\left\{(\widehat{f}_n^i - \widetilde{f}_n^i)^2\right\} = E\left\{[(\widehat{f}_n^i - (1 - p)(\widetilde{f}_{ref}^j + \widehat{r}_n^i) - p\widetilde{f}_{n-1}^k)]^2\right\} \\ &= (1 - p)E\left\{(\widehat{f}_n^i - \widehat{r}_n^i - \widetilde{f}_{ref}^j)^2\right\} + pE\left\{(\widehat{f}_n^i - \widetilde{f}_{n-1}^k)^2\right\} \\ &\approx (1 - p)E\left\{(\widehat{f}_{ref}^j - \widetilde{f}_{ref}^j)^2\right\} + pE\left\{(\widehat{f}_n^i - \widetilde{f}_{n-1}^k)^2\right\} + pE\left\{(\widetilde{f}_{n-1}^k - \widetilde{f}_{n-1}^k)^2\right\} \\ &= (1 - p)d_{ep_ref} + p(d_{ec_rec} + d_{ec_ep}), \end{aligned} \quad (3)$$

The latter term composed of d_{ec_rec} and d_{ec_ep} denotes the error-concealed distortion that would propagate to the following frames. In detail, d_{ec_rec} is the newly

incurred distortion between the error concealed pixel and the reconstructed pixel at the encoder. And if the pixel is error concealed, then the error-propagated distortion of the error concealed pixel d_{ec-ep} is inherited. In order to obtain the end-to-end distortion, the error concealment distortion for pixel i in frame n is computed by:

$$\begin{aligned}
 d_{ec} &= E\left\{(f_n^i - \tilde{f}_{n-1}^k)^2\right\} \\
 &\approx E\left\{(f_n^i - \hat{f}_n^i)^2\right\} + E\left\{(\hat{f}_n^i - \tilde{f}_{n-1}^k)^2\right\} \\
 &= d_s + d_{ec-rec} + d_{ec-ep}.
 \end{aligned} \tag{4}$$

Combining (2), (3) and (4), the end-to-end distortion can be computed by the linear combination of the source distortion d_s and the error-propagated distortion d_{ep} . Note that if sub-pel motion compensation is used, the reference sample could point to a sub-pel position. Further considering the use of in-loop filter, the computation of D_{ep-ref} becomes more complex. To tackle these problems, we propose a block-based algorithm to track the potential distortion.

2.2 Block-Based Potential Distortion Tracking

Assume a distortion map D_{ep} is defined for each frame on a block basis (e.g. 4x4). The first frame in a sequence is coded with intra mode without considering the error propagation. Then the distortion map of the first frame is derived for coding the subsequent frames. For coding the other frames, the distortion maps of their previous frames indicating the possible influence of propagated errors are referenced to select the proper coding modes. After coding each frame, the associated distortion map is derived accordingly. The referenced error-propagation distortion of the k th block in the m th MB of the n th frame, denoted as $D_{ep-ref}(n, m, k)$, is computed by weighting the potential error-propagated distortion of the surrounding blocks in the reference frames that overlap with the motion-compensated blocks. Since there is not reference frame for the intra-coded MB, the propagated errors from the previous frames are suppressed and therefore $D_{ep-ref}(n, m, k)$ in terms of the intra mode equals to zero.

After the current frame is coded, the distortion map D_{ep} in terms of the current frame is then derived according to (3). The calculation of $D_{ec-ep}(n, m, k)$ and $D_{ec-rec}(n, m, k)$ depends on the employed error concealment method at the decoder side. In the H.264 non-normative decoder, the lost MB is reconstructed by copying some blocks in the previous frames. In this case, $D_{ec-ep}(n, m, k)$ can be derived from the potential error-propagated distortions of these blocks. The distortion map is stored for the following frame encoding. Notice that for coding the first frame, the referenced distortion maps are null, which indicates that the first frame does not suffer from the propagated errors. Moreover, the distortion map of the first frame is only associated with error-concealed distortion D_{ec-rec} . The end-to-end distortion of the k th block in the m th MB of the n th frame is computed as the sum of source distortion and the error-propagated distortion.

2.3 Simulations and Discussion

To test the performance of the proposed distortion estimation scheme, we simulate packet loss in H.264 video coding and use this scheme to estimate the end-to-end distortion. Each row of macroblocks is transmitted in a separate packet. If a MB is lost, the decoder copies the co-located MB in the previous frame. This simple concealment method is used in both encoder and decoder. The mode selection is performed according to the ER-RDO algorithm. Only the first frame is encoded as I frame, the left frames are encoded as P frame. The number of reference frames is one. The Foreman QCIF video sequence is simulated 30,100,500 times and the average frame-level decoder distortion is computed. The estimation result is shown in Fig. 1. The accuracy of experimental results suggests that the proposed distortion model is consistent to the statistical K-time decoding at the encoder side, however, the added complexity is much lower than the latter.

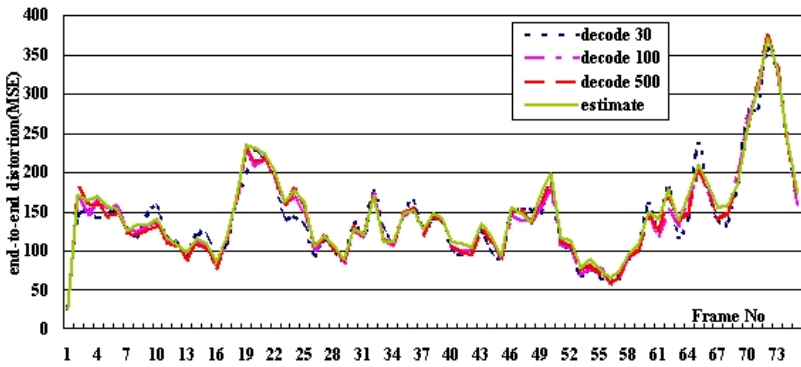


Fig.1. end-to-end distortion estimation results for Foreman sequence, r=64kbps, packet loss rate=10%

3 RDO-Based Error Resilient Coding

3.1 RDO-Based Error Control

Suppose O denotes the set of all selectable coding options in terms of an MB. The coding option o^* of the m th MB in the n th frame is selected to be the one that minimizes the cost given by:

$$J(n, m, o) = D(n, m, o) + \lambda R(n, m, o) \tag{5}$$

According to the previous section, since the potential channel distortion is also considered in the expected overall distortion, the relation between the rate and the overall distortion changes as well. Accordingly, the Lagrange parameter λ

should be properly selected. Similar to the Lagrange parameter selection scheme for the error-free environment in [10], we derive the new Lagrange parameter as $(1 - p)\lambda$, where λ is the Lagrange parameter in the error-free environment. Then, the proposed RDO-based error-resilient coding for H.264 encoder in the packet-loss environment is performed as follows. Since the channel distortion of the B frame would not propagate to the following P frames, it is unnecessary to define the distortion map for the B frame to store the potential error-propagated distortion. The coding mode selection of B frames can be the same as that in the P frame coding. For simplicity, we assume that B frames are not used. In H.264, the coding mode of P frames is selected to be one of the 2 intra modes and 8 inter modes. In terms of inter mode, the reference can be from one of the several previous frames. We assume a simple error concealment scheme at the decoder side is used. If a MB is lost, the decoder simply copies the co-located MB in the previous decoded frame.

Notice that the distortions introduced while the current MB is lost are independent of the coding option. Supposing the packet loss rate p is known at the encoder side, according to the (2) and (5), the coding option can be selected with

$$\begin{aligned} o^*(n, m) = \underset{o \in O}{\operatorname{argmin}} & ((1 - p)(D_s(n, m, o) + D_{ep_ref}(n, m, o)) + pD_{ec}(n, m) \\ & + (1 - p)\lambda R) = \underset{o \in O}{\operatorname{argmin}} (D_s(n, m, o) + D_{ep_ref}(n, m, o) + \lambda R). \end{aligned} \quad (6)$$

After the current frame is encoded, the distortion map is derived according to the (3) for the encoding of the future frames.

3.2 Discussion of Rate Control

While we investigate the error resilience feature of the proposed R-D based model in H.264/AVC, we also analyze the effect of the rate control scheme adopted in the H.264/AVC test model [11]. Basically, rate control resolves two main problems, i.e. bit allocation and quantization parameter adjustment. In H.264 test model, it consists of three tightly consecutive components: GOP level rate control, picture rate control and the optional basic unit level rate control. The basic unit is defined as a group of successive MBs in the same frame. For detailed algorithm of the basic unit level rate control, we refer to [11].

Supposing a slice/package is taken as a basic unit, rate control can be easily jointly implemented with the coding mode selection scheme described in the previous subsection. The remained problem is whether or not the accuracy and coding efficiency of rate control can satisfy the requirements. Obviously, the rate allocation at GOP level does not have effect. The R-D model in picture/basic unit layer is:

$$R = C_1 \times \frac{MAD}{Q_{step}} + C_2 \times \frac{MAD}{Q_{step}^2} - M, \quad (7)$$

where M is the total number of header bits and motion vector bits, and C_1 and C_2 are two adaptively adjusted coefficients. The key point is about how

to estimate the *MAD* prior to doing the current basic unit rate control. In the traditional RDO-based framework, it is proven that the current *MAD* can be estimated with the linear regression model using the actual *MAD* of the previous picture or the co-located basic units in previous picture. In the proposed coding mode selection scheme, this *MAD*-based linear regression method is still applicable because the same mode selection scheme is used for each picture and the statistics of *MAD* remains very similar. Simulations have shown that the basic unit level rate control (i.e. taking the slice as the basic unit) has the very similar coding efficiency to the fixed quantization parameter coding.

3.3 Experimental Results

Some experiments have been carried out to verify the performance of the proposed algorithm. Two algorithms are compared: the proposed coding mode decision algorithm and ER-RDO. The testing platform is the H.264 reference software JM7.5c [11]. In the default ER-RDO algorithm, $K=500$ decoders are simulated in the encoder. Only the first frame is encoded as I frame, and the following frames are encoded as P frames. Five coded sequences are generated for each algorithm: Foreman@64kbps (QCIF, 7.5fps, and denoted as Fore_64k), Foreman@144kbps (QCIF, 7.5fps, and denoted as Fore_144k), Hall Monitor@32kbps (QCIF, 10fps, and denoted as Hall_32k), Paris@144kbps (CIF, 15fps, and denoted as Paris_144k) and Paris@384kbps (CIF, 15fps, and denoted as Paris_384).

The packet loss situation is simulated according to the error resilience testing conditions specified in [12]. The coded sequences were decoded after packet loss simulation under packet loss rates 3, 5, 10 and 20%. Note that there are 4 packets per frame for QCIF and 9 packets for CIF. The 40 bytes of IP/UDP/RTP headers per packet have been taken into account. The simple previous frame copy error concealment method is used in all simulations. The average YPSNR values of the coded sequences except the first frame under different packet loss rate are shown in Table 1. The results show that the proposed algorithm outperforms ER-RDO in terms of transmission efficiency in all cases. Moreover, the most significant difference lies in the computing complexity. In these experiments, the running time of ER-RDO is about as 25 times long as the original algorithm without error control, whereas the running time of the proposed algorithm is very similar to the original algorithm.

Table 1. Comparison results of average PSNR (in dB) at different packet loss rates

Sequences	LossRate: 3%		LossRate: 5%		LossRate: 10%		LossRate: 20%	
	Proposed	ER-RDO	Proposed	ER-RDO	Proposed	ER-RDO	Proposed	ER-RDO
Fore_64k	30.31	30.21	29.48	29.42	27.60	27.46	25.58	25.50
Fore_144k	34.36	34.23	33.22	33.15	30.85	30.60	28.17	28.11
Hall_32k	33.58	33.25	33.44	33.09	32.29	31.87	31.19	30.93
Paris_144k	27.51	26.71	27.01	26.15	25.93	25.59	24.84	24.54
Paris_384k	33.08	32.68	32.29	31.72	33.74	33.34	29.33	28.98

4 Conclusion

An optimized rate-distortion model for H.264 video encoder in the packet loss environment has been presented in this paper. The encoder keeps tracking the distortion on a block basis while taking into account the source characteristics, network conditions as well as the error concealment method. The proposed model reveals the inherent relationship between the potential error-propagated distortion and the characteristics of the input source video data. Compared to the error robust rate-distortion optimization method in H.264 test model, the proposed model performs better in terms of both transmission efficiency and computational complexity. Furthermore, although this algorithm is proposed for H.264 encoder, it is also feasible to be used in other standard-compliant video encoders.

References

1. Stuhlmüller, M., Farber, N., Link, N., Girod, B.: Analysis of Video Transmission over Lossy Channels. *IEEE J. Selected Areas in Communications*, Vol. 18. 6 (2000) 1012-1032
2. Wang, Y., Zhu, Q. F.: Error Control and Concealment for Video Communication: A Review. *Proc. IEEE*, Vol. 86. 5 (1998) 974-997
3. Cote, G., Kossentini, F.: Optimal Intra Coding of Blocks for Robust Video Communication over the Internet. *Image Commun.* 9 (1999) 25-34
4. Zhu, Q. F., Kerofsky, L.: Joint Source Coding, Transport Processing and Error Concealment for H.323-based Packet Video. *Proc. SPIE, VCIP'99*, Vol. 3653. San Jose, CA (1999) 52-62
5. Haskell, P., Messerschmitt, D.: Resynchronization of Motion-Compensated Video Affected by ATM Cell Loss. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 3. (1992) 545-548
6. He, Z. H., Cai, J. F., Chen, C. W.: Joint Source Channel Rate-Distortion Analysis for Adaptive Mode Selection and Rate Control in Wireless Video Coding. *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 12. 6 (2002) 511-523
7. Zhang, R., Regunathan, S. L., Rose, K.: Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience. *IEEE J. Selected Areas in Communications*, Vol. 18. 6 (2000) 966-976
8. Stockhammer, T., Kontopodis, D., Wiegand, T.: Rate-Distortion Optimization for JVT/H.26L Coding in Packet Loss Environment. *Proc. PVW. Pittsburgh, PY* (2002)
9. ITU-R Rec. H.264 — ISO/IEC 14496-10 AVC: Draft Text. Joint Video Team document JVT-E146D37 (2002)
10. Wiegand, T., Girod, B.: Lagrangian Multiplier Selection in Hybrid Video Coder Control. *Proc. ICIP2001. Thessaloniki, Greece* (2001)
11. <http://bs.hhi.de/~suehring/tml/download/jm75c.zip>
12. Wenger, S.: Common Conditions for Wire-line, Low Delay IP/UDP/RTP Packet Loss Resilient Testing. ITU-T VCEG document VCEG-N79r1 (2001)

Enhanced Stochastic Bit Reshuffling for Fine Granular Scalable Video Coding

Wen-Hsiao Peng, Tihao Chiang, Hsueh-Ming Hang, and Chen-Yi Lee

National Chiao-Tung University
1001 Ta-Hsueh Rd., HsinChu 30010, Taiwan.
pawm@mail.sisl2lab.org, {tchiang, hmhang}@mail.nctu.edu.tw,
cylee@cc.nctu.edu.tw

Abstract. In this paper, we propose an enhanced stochastic bit reshuffling (SBR) scheme to deliver better subjective quality for fine granular scalable (FGS) video coding. Traditional bit-plane coding in FGS algorithm suffers from poor subjective quality due to zigzag and raster scanning order. To tackle this problem, our SBR rearranges the transmission order of each bit by its estimated rate-distortion performance. Particularly, we model the transform coefficient with a maximum likelihood based Laplacian distribution and incorporate it into the context probability model for content-aware parameter estimation. Moreover, we use a dynamic priority management scheme for the SBR. Experimental results show that our enhanced SBR together with context adaptive binary arithmetic coding offers up to 1.5dB PSNR improvement and shows better visual quality as compared to the scheme in MPEG-4 FGS.

1 Introduction

MPEG-4 fine granularity scalability (FGS) [1] offers a bit-plane coding scheme using variable length code (VLC). The same approach is also widely used in most of the advanced FGS algorithms for fine granular SNR scalability. For each bit-plane, current approach performs the coding in a block-by-block manner. When the enhancement-layer is partially decoded, zigzag and raster scanning order may only refine the upper part with one extra bit-plane. Such uneven quality distribution causes subjective quality degradation.

In our prior work [2], a context adaptive binary arithmetic coding (CABAC) with an in-bit-plane bit reshuffling scheme was proposed to deliver higher coding efficiency and better subjective quality for the FGS algorithms. To improve coding efficiency, we partition each transform coefficient into significant and refinement bits. For the significant bit, we construct context model based on both the energy distribution in a block and the spatial correlation in the adjacent blocks. The spatial correlation is considered using the significance status of the co-located coefficients in the adjacent blocks as context model. For the refinement bit, we simply use a fixed context probability model for coding. To improve subjective quality, the in-bit-plane reshuffling scheme reorders the coding bits of

a bit-plane in a rate-distortion sense. Particularly, for each bit, we use the associated non-zero context probability model to approximate its distortion reduction at the decoder and exploit the binary entropy to estimate its bit rate. Experimental results in [2] show that our CABAC significantly improves the coding efficiency while the in-bit-plane reshuffling only shows slight improvement on the subjective quality.

In this paper, we propose an enhanced stochastic bit reshuffling (SBR) scheme based on the CABAC in [2]. Instead of bit-plane by bit-plane coding, our enhanced SBR allows different coefficients be coded with different bin numbers. Moreover, for more accurate rate-distortion function estimation, we exploit a maximum likelihood based discrete Laplacian distribution with the context probability model to approximate the transform coefficients. Furthermore, to implement the SBR, we propose a dynamic priority management scheme that makes the SBR content aware considering the content dependent priority and a rate-distortion data update mechanism. Experimental results show that our CABAC [2] with the enhanced SBR can offer up to 1.5dB PSNR improvement and show better subjective quality as compared to the traditional scheme in MPEG-4 FGS.

The remainder of this paper is organized as follows: Section 2 introduces the SBR. Section 3 shows our parameter estimation schemes. Section 4 presents the dynamic coding flow for the SBR. Section 5 shows the experimental results. Finally, Section 6 summarizes and concludes our work in this paper.

2 Stochastic Bit Reshuffling (SBR)

To address the uneven subjective quality issue, we propose a content-aware SBR scheme. Instead of deterministic zigzag and raster scanning order, we propose to reshuffle the coding bits in an optimized stochastic rate-distortion sense.

The bit reshuffling concept and criterion was first proposed in [3] for improving the rate-distortion performance of wavelet based image codec. For the reshuffling, in [3] each input bit is first assigned with two factors that are squared error reduction, ΔD , and coding cost, ΔR . With these parameters, the bit reshuffling reorders the input bits such that the associated $(\Delta D/\Delta R)$ is in descending order. Such an order leads to minimum distortion at any bit rate. However, the decoder generally does not have actual ΔD and ΔR for each input bit. Thus, estimated ΔD and ΔR are used to avoid sending the coding order. With the same estimation scheme at both encoder and decoder, the coding order can be implicitly known to both sides. Eq. (1) shows the condition of optimal coding order in *stochastic* rate-distortion sense where \hat{E} denotes taking estimation, the subscript of ΔD and ΔR represents the bit identification and the one of $(\hat{E}[\Delta D]/\hat{E}[\Delta R])$ specifies the coding order.

In this paper, we use the same concept for the enhancement-layer bit reshuffling. Particularly, our estimation incorporates the context probability model to make the priority assignment content-aware so that the subjective quality can be improved. Moreover, we follow the optimized coding order in stochastic

rate-distortion sense as in Eq. (1) to have similar or even better rate-distortion performance.

$$\left(\frac{\widehat{E}[\Delta D_i]}{\widehat{E}[\Delta R_i]}\right)_1 \geq \left(\frac{\widehat{E}[\Delta D_j]}{\widehat{E}[\Delta R_j]}\right)_2 \geq \left(\frac{\widehat{E}[\Delta D_k]}{\widehat{E}[\Delta R_k]}\right)_3 \geq \dots \geq \left(\frac{\widehat{E}[\Delta D_l]}{\widehat{E}[\Delta R_l]}\right)_N \quad (1)$$

3 Parameter Estimation

To estimate the ΔD and ΔR of each bit, we use their expectations. For calculating the expectation, we need the probability distribution of each transform coefficient. However, decoder generally does not have actual distribution. Thus, to minimize the overhead, we model each 4x4 integer transform coefficient [4] with discrete Laplacian distribution as defined in Eq. (2) where $X_{n,k}$ represents the n th zigzag ordered coefficient of block k and $x_{n,k}$ stands for its outcome. Particularly, we assume that the co-located coefficients are independently and identically distributed (i.i.d.).

$$P[X_{n,k} = x_{n,k}] \triangleq \frac{1 - \alpha_n}{1 + \alpha_n} \times (\alpha_n)^{|x_{n,k}|} \text{ where } n = 0 \sim 15 \quad (2)$$

3.1 Estimation of Laplacian Parameter α_n

In Eq. (2), α_n ($n:0 \sim 15$) is to be estimated. For the estimation, we use maximum likelihood principle. Given a set of M observed data and a presumed joint probability with an unknown parameter, the maximum likelihood estimator for the unknown parameter is the one that maximizes the joint probability.

For an enhancement-layer frame with M 4x4 blocks, the joint probability of the co-located coefficients at n th zigzag position can be written as in Eq. (3). According to the i.i.d. assumption, we can simplify the joint probability as M multiplication terms. Further, by substituting Eq. (2) into Eq. (3), we can obtain a close form formula for the joint probability.

$$\begin{aligned} &P[X_{n,1} = x_{n,1}, X_{n,2} = x_{n,2}, \dots, X_{n,k} = x_{n,k}, \dots, X_{n,M} = x_{n,M}] \\ &= \left(\frac{1 - \alpha_n}{1 + \alpha_n}\right)^M \times (\alpha_n)^{\sum_{k=1}^M |x_{n,k}|} \end{aligned} \quad (3)$$

By definition, the maximum likelihood estimator of α_n is the one that maximizes Eq. (3). To find the solution, one can take differentiation with respect to α_n and look for the root. Eq. (4) shows the maximum likelihood estimator of α_n based on Eq. (3). Note that the estimation is conducted at the encoder and the estimated parameters are transmitted to the decoder.

$$\alpha_n = \frac{-\mu_x^{-1} + \sqrt{(\mu_x^{-1})^2 + 4}}{2} \text{ where } \mu_x = \frac{\sum_{k=1}^M |x_{n,k}|}{M} \quad (4)$$

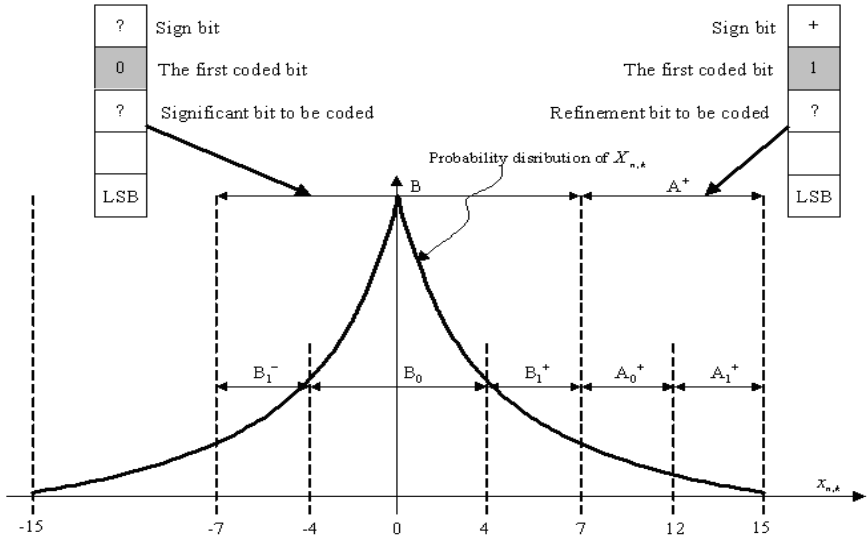


Fig. 1. Examples of ΔD estimation for significant bit and refinement bit

3.2 Estimation of ΔD

To estimate the ΔD of a coefficient bit, we calculate the expected squared error reduction at the decoder. Since the decoding of a coefficient bit is to reduce the uncertainty of a coefficient, we can calculate the expected squared error reduction from the decrease of uncertainty interval.

In Fig. 1, we depict the estimated distribution of a 4-bit coefficient and give two examples for illustrating the decrease of uncertainty interval. Without loss of generality, we show the example of significant bit at the left hand side where the first bit is coded as zero. On the other hand, the right hand side depicts the case of refinement bit where the first bit is non-zero. From the coded bits, we can identify the uncertainty interval in which the actual value is located. For the significant bit case, we know that the actual value is confined within the interval B . Similarly, for the refinement bit case, we learn that the actual value falls in the interval A^+ . Given the interval derived from the previously coded bits, the next bit for coding is to further decrease the uncertainty interval. For instance, the significant bit to be coded is to determine that the actual value is in the subinterval B_0 or $\{B_1^+ \cup B_1^-\}$. Similarly, the refinement bit to be coded is to determine that the actual value is in the subinterval A_0^+ or A_1^+ .

From the decrease of uncertainty interval, we can calculate the reduction of expected squared error. Specifically, at the decoder, the expected squared error in an interval is the variance within the interval. Thus, we can express our ΔD estimation as the reduction of variance. Eq. (5) formulates our ΔD estimation for the significant bit case in Fig. 1 where $\text{Var}[X_{n,k}|X_{n,k} \in B]$ represents the

conditional variance of $X_{n,k}$ given that $X_{n,k}$ is in the interval B . Similarly, we have the conditional variances for the other subintervals. Since we do not know in which subinterval the actual value is located, each subinterval variance is further weighted by its probability. To simplify the expression, we merge the variances of B_1^+ and B_1^- in Eq. (5) because they are identical.

$$\begin{aligned}
& \widehat{E}[\Delta D_{n,k,B,significant}] \\
& \triangleq \text{Var}[X_{n,k}|X_{n,k} \in B] \\
& - (P[X_{n,k} \in \{B_1^+ \cup B_1^-\} | X_{n,k} \in B]) \times \text{Var}[X_{n,k}|X_{n,k} \in B_1^+] \\
& - P[X_{n,k} \in B_0 | X_{n,k} \in B] \times \text{Var}[X_{n,k}|X_{n,k} \in B_0] \\
& \cong \text{Var}[X_{n,k}|X_{n,k} \in B] \\
& - \text{SignificantContext}P(\text{ContextIndex}(n,k,B), 1) \times \text{Var}[X_{n,k}|X_{n,k} \in B_1^+] \\
& - \text{SignificantContext}P(\text{ContextIndex}(n,k,B), 0) \times \text{Var}[X_{n,k}|X_{n,k} \in B_0]
\end{aligned} \tag{5}$$

In Eq. (5), the co-located significant bits within the same interval have the same estimated ΔD because the co-located coefficients have identical Laplacian model. To perform the reshuffling with content aware so that the regions that contain more energy are with higher coding priority, in Eq. (5) the subinterval probabilities are approximated with the context probability models where $\text{SignificantContext}P(\text{ContextIndex}(n,k,B), 1)$ denotes the non-zero context probability model for the significant bit of $X_{n,k}$ in the interval B and the $\text{SignificantContext}P(\text{ContextIndex}(n,k,B), 0)$ represents its zero probability. Recall that we exploit the significance status of co-located coefficients in the adjacent blocks as significant bit context model [2]. Using the context probability model for substitution makes the ΔD estimation of significant bit become region dependent, i.e., content aware.

Following the same procedure, one can obtain the ΔD estimation for the other significant bits and refinement bits. Eq. (6) shows the one for the refinement bit example in Fig. 1. Particularly, the conditional probability in Eq. (6) is from the estimated Laplacian model.

$$\begin{aligned}
& \widehat{E}[\Delta D_{n,k,A^+,refinement}] \\
& \triangleq \text{Var}[X_{n,k}|X_{n,k} \in A^+] \\
& - P(X_{n,k} \in A_1^+ | X_{n,k} \in A^+) \times \text{Var}[X_{n,k}|X_{n,k} \in A_1^+] \\
& - P(X_{n,k} \in A_0^+ | X_{n,k} \in A^+) \times \text{Var}[X_{n,k}|X_{n,k} \in A_0^+]
\end{aligned} \tag{6}$$

3.3 Estimation of ΔR

To estimate the expected coding bit rate of an input bit, we use the binary entropy¹ which represents the minimum expected coding bit rate for an input bit. Eq. (7) defines our ΔR estimation for the significant bit example in Fig.

¹ $H_b(P(1)) = -P(1) \times \log_2(P(1)) - (1 - P(1)) \times \log_2(1 - P(1))$, where $P(1)$ is the non-zero probability of the coding bit

1. The first term represents the binary entropy of a significant bit using the associated context probability model as an argument while the second term denotes the cost from a sign bit. The sign bit is considered as partial cost of significant bit because the decoder can only perform the reconstruction after the sign is received. Recall that each sign bit averagely consumes one bit and it is only coded after a non-zero significant bit. Thus, the cost of sign is weighted by the non-zero context probability model.

$$\begin{aligned} & \widehat{E}[\Delta R_{n,k,B,significant}] \\ & \triangleq H_b(\text{SignificantContextP}(\text{ContextIndex}(n,k,B),1)) \\ & + \text{SignificantContextP}(\text{ContextIndex}(n,k,B),1) \times 1 \end{aligned} \quad (7)$$

Eq. (8) illustrates our ΔR estimation for the refinement bit example in Fig. 1. For the binary entropy calculation, we use the estimated probability of refinement bit as an argument. For instance, we use $P(X_{n,k} \in A_1^+ | X_{n,k} \in A^+)$ as the argument for the refinement bit example in Fig. 1. For the other significant bits and refinement bits, one can use the same methodology to estimate the ΔR .

$$\begin{aligned} & \widehat{E}[\Delta R_{n,k,A^+,refinement}] \\ & \triangleq H_b(P(X_{n,k} \in A_1^+ | X_{n,k} \in A^+)) \end{aligned} \quad (8)$$

4 Dynamic Priority Management for SBR

To maintain the coding priority, we exploit two dynamic coding lists for the reshuffling of significant and refinement bits. Each bit in the associated list is allocated a register to record its bit location and estimated rate-distortion data. Particularly, to avoid sending the bin position for each bit, the coding of a coefficient is sequentially from the MSB to the LSB. In addition, to avoid coding the redundant bits after the EOSP, the coding of significant bits in Part II always follows zigzag order. In [2], EOSP denotes the location of last non-zero significant bit of a bit-plane. For each bit-plane, the significant bits in Part II refer to those after the EOSP of previously coded bit-plane in zigzag order. Given these constraints, our priority management includes the following 5 steps:

1. **Coding list initialization:** The initialization estimates the rate-distortion data for all the DC coefficient bits at the MSB bit-plane of the enhancement-layer frame and puts the data in the coding lists.
2. **Coding list reshuffling:** After the initialization, we perform the reshuffling to identify the highest priority bit in the lists, i.e., the one with maximal $(\widehat{E}[\Delta D] / \widehat{E}[\Delta R])$. In addition, the reshuffling is performed after the coding of each input bit.
3. **CABAC:** Once the highest priority bit is identified, we follow the CABAC scheme in [2] for coding.

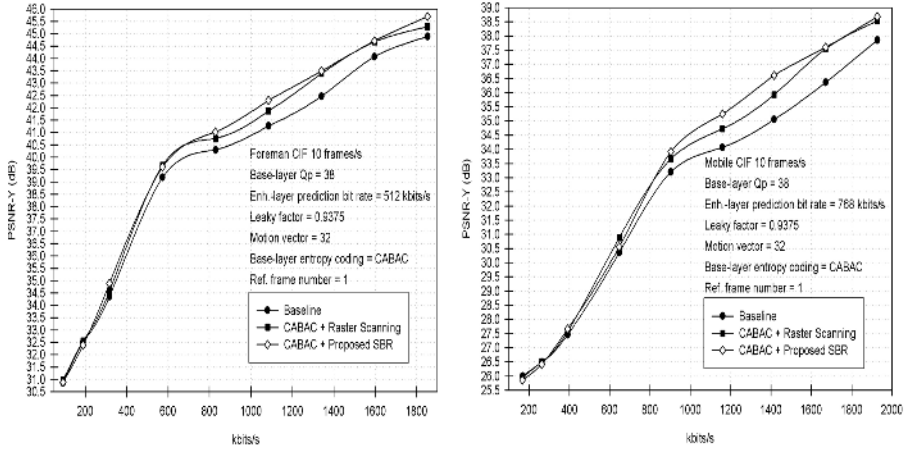


Fig. 2. PSNR comparison of traditional bit-plane coding (Baseline), CABAC [2] + Raster Scanning and CABAC + Proposed SBR.



Fig. 3. Subjective quality comparison of Baseline and CABAC [2] + SBR with enh.-layer truncated at 384kbit/s.

- 4. Rate-distortion data update:** After the coding of each input bit, we update parts of the rate-distortion data in the coding lists. Such update is to fully utilize the coded information to enhance the content aware bit reshuffling. Specifically, we update those registers whose context index or context probability model are changed after the coding of highest priority bit. Particularly, the rate-distortion data of refinement bit is not required for update since it is from the fixed Laplacian model.

5. **Input bit inclusion:** After the update, we further include the adjacent bit in the lower bit-plane and the next zigzag ordered coefficient bit in the same bit-plane for reshuffling. The steps 2 to 5 are repeated until all the input bits are coded.

5 Experiments

In this Section, we assess the rate-distortion performance of our SBR objectively and subjectively. For the experiments, we use H.264 JM4 [4] as base-layer and RFGS [5] as enhancement-layer prediction scheme. For comparison, different bit-plane coding schemes use identical encoder parameters. Particularly, the scheme in MPEG-4 FGS is used as baseline.

Fig. 2 shows that the CABAC [2] with enhanced SBR further boosts the PSNR by 0.2~0.5dB as compared to one with raster scanning. Moreover, it reveals 1.0~1.5dB improvement over the baseline. In addition, Fig. 3 shows that the enhanced SBR offers better subjective quality. The baseline reveals obvious blocking artifact at the lower part of the decoded frame. In contrast, our CABAC with enhanced SBR shows more uniform quality over the entire frame.

6 Conclusion

In this paper, we present an enhanced SBR scheme to improve the subjective quality of traditional bit-plane coding in FGS algorithms. We generalize the concept of bit-plane coding and reshuffle each coding bit according to its estimated rate-distortion performance. Experimental results show that our CABAC [2] with enhanced SBR can deliver higher coding efficiency and better subjective quality. Furthermore, with appropriate modification, the proposed SBR can be applied in other embedded entropy coding.

References

1. W. Li, "Overview of Fine Granularity Scalability in MPEG-4 Standard," *IEEE Trans. on Circuits Syst. for Video Technol.*, vol. 11, no. 3, pp. 301–317, 2001.
2. W. H. Peng, C. N. Wang, T. Chiang, and H. M. Hang, "Context Adaptive Binary Arithmetic Coding With Stochastic Bit Reshuffling For Advanced Fine Granularity Scalability," in *IEEE ISCAS*, 2004.
3. J. Li and S. Lei, "An Embedded Still Image Coder with Rate-Distortion Optimization," *IEEE Trans. on Image Processing*, vol. 9, no. 7, pp. 297–307, 2003.
4. T. Weigand, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC)," *JVT-G050*, 2003.
5. H. C. Huang and T. Chiang, "Stack Roubst Fine Granularity Scalability," in *IEEE ISCAS*, 2004.

High-Performance Motion-JPEG2000 Encoder Using Overlapped Block Transferring and Pipelined Processing

Byeong-Doo Choi¹, Min-Cheol Hwang¹, Ju-Hun Nam¹,
Kyung-Hoon Lee², and Sung-Jea Ko¹

¹ Department of Electronics Engineering, Korea University
5-1 Anam-Dong, Sungbuk-ku, Seoul 136-701, Korea
sjko@dali.korea.ac.kr

² Electronics and Telecommunications Research Institute

Abstract. This paper presents effective DSP implementation strategies for real-time JPEG2000 encoder system, an overlapped block transferring (OBT) for DWT and a pipelined processing of passes (PPP) for EBCOT Tier-1. The proposed OBT method can significantly improve the performance of the lifting algorithm for DWT by increasing the cache hit rate. Moreover the PPP method reduces the processing time of EBCOT Tier-1 by pipelined processing of the three coding passes of the same bit-plane. Experimental results show that Motion-JPEG2000 DSP system meets the common requirement of the real-time video coding [30 frames/s (fps)] and is proven to be a practical and efficient DSP solution.

Keywords: JPEG2000, OBT, DSP, Wavelet.

1 Introduction

JPEG2000 compression standard has been created to provide high compression efficiency compared to JPEG [1]. It includes a rich set of features such as improved compression efficiency, lossy to lossless compression, multiple resolution representation, embedded bit-stream, region-of-interest (ROI) coding, and error resilience [2,3]. The major difference between previously proposed wavelet-based image compression algorithms such as EZW or SPIHT is that EBCOT on JPEG2000 operates on independent, non-overlapping blocks which are coded in several bit layers to create an embedded, scalable bitstream. This property gives a possibility to be able to increase the operating performance.

In this paper, we propose the overlapped block transferring (OBT) method, based on the cache performance to improve DWT. Instead of the line-based lifting scheme, an image is divided into overlapped subblocks and then each overlapped subblock is processed by a 2-D lifting algorithm to increase the cache hit rate. We show that the OBT-based lifting scheme can increase the performance of the DWT drastically. Moreover, we propose a pipelined processing of passes

(PPP) for fast implementation of EBCOT Tier-1. This method reduces the processing time of EBCOT Tier-1 by processing the three coding passes of the same bit-plane in parallel.

The paper is organized as follows: The proposed architecture of Motion-JPEG2000 is introduced in Section 2. The OBT-based lifting scheme is proposed in Section 3, and pipeline processing of passes for EBCOT is proposed in Section 4. In Section 5, the performance of the proposed system is discussed and conclusions are given in Section 6.

2 DSP Implementation of Motion-JPEG2000

Fig. 1 shows the block diagram for the hardware architecture of the implemented Motion-JPEG2000 system. This system consists of three modules, which are main processor (DSP : TMS320C6416), video interface module and external interface module. External interface module is designed to interface with communication protocol.

Fig. 2 shows the encoding steps of JPEG2000 coding system. For encoding, an image is split into rectangular structures called tiles. The tiles are encoded independently as if they are different images. The DWT is applied on a component in the tile to decompose it into a number of wavelet subbands at different levels and resolutions. In the next section, we show that the proposed OBT method can significantly improve the performance of the lifting algorithm for DWT by increasing the cache hit rate. The quantized subbands are divided into

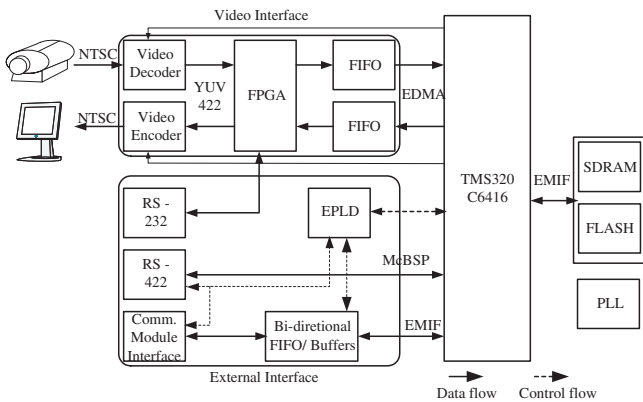


Fig. 1. Block diagram of the implemented hardware platform.

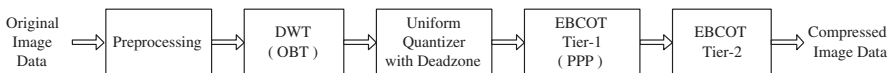


Fig. 2. The encoding steps of JPEG2000 coding system.

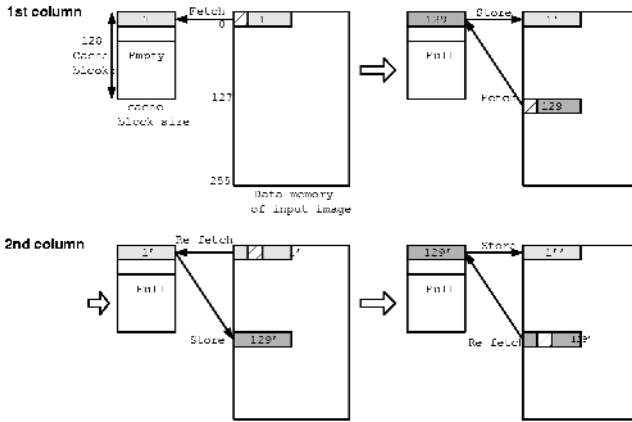


Fig. 3. Cache-misses in vertical wavelet filtering.

codeblocks. The codeblocks are entropy coded along the bit plane using the EBCOT. Moreover, we also show that the proposed PPP method reduces the processing time of EBCOT Tier-1 by processing the three coding passes of the same bit-plane in parallel. Since this scheme depends on a per-block quad-tree structure, the independent code-blocks are passed down the coding pipeline and generate separate bitstreams.

3 OBT-Based Lifting Scheme for Efficient Cache Utilization

The lifting algorithm is a fast computing technique of the DWT. However, in a point of view of memory management, it still has severe cache-miss problems during the execution of the vertical wavelet filtering. A number of cache-misses make the processing time increase critically. Thus, even though the lifting algorithm requires few execution of CPU, the processing time of DWT cannot be reduced remarkably without the memory management to reduce cache-miss. Fig. 3 shows the reason why the cache-misses occur in the vertical filtering of DWT. When filtering the first column, memory blocks are fetched as many as the height of image. In the second column, memory blocks, which were fetched in the filtering of the first column, must be re-fetched because cache blocks of upper-side data of an image are written to external memory while down-side data is loaded in case that image height is larger than the number of cache blocks. These re-fetches of same data memory are iterated at every column. As a consequence, this means that the vertical filtering of the conventional lifting scheme lets same data memory be loaded on cache as many as the data cache block size. The data cache miss rate is drastically increased.

This problem can be improved by partitioning an entire image to blocks. Conventionally, in order to perform a lifting scheme, the image rows are filtered

in the horizontal direction, and the image columns are filtered in the vertical direction. However, our approach partitions an entire image into blocks to fit into the cache size and reorders the processing sequence to be processed block by block. It reduces the cache miss rate because data, which is fetched in horizontal filtering, is remained on cache until the vertical filtering of current block is completed. Thus, a whole image data can not be loaded on data cache.

This method is seemed to remove perfectly a cache-miss problem. However, two problems exist in the proposed method. One is that coefficients of edge cannot be filtered independently without coefficients of adjacent blocks. The other is that the data, which is not aligned by cache block size, must be fetched two times. This means that the redundancy of re-fetch exists yet.

In order to solve these two problems, we propose an overlapped block transferring (OBT) method. This method is based on hierarchical memory architecture. The memory architecture of JPEG2000 encoding system is composed of two layers of data caches (L1D, L2D) and an external memory.

The principle of the OBT method is that DMA operating independently without CPU execution transfers the image data to L2D from the external memory by block size equal to the cache size. Unlike the external memory, the address of the data memory on L2D is aligned by cache block size. Since the L2D can not hold a whole image of large size, DMA transfers data blocks from the external memory to L2D and from L2D to the external memory repeatedly like double-buffering. Fig. 4 shows this mechanism. The data of the first column of blocks is transferred to L2D. After the first column of blocks is processed, this data is moved to the external memory and the next column of blocks is transferred to same location of L2D by DMA.

Fig. 5 shows that the adjacent blocks are overlapped with each other along the horizontal direction. Area 1 in light gray is completely wavelet processed, whereas Area 2 in dark gray contains data lifted partially. Thus, the next block for the 2-D lifting is placed to include Area 2 as well as Area 3 in light gray. The remaining horizontal lifting steps for pixel values in Area 2 are completed, and then the 2-D lifting scheme is processed for Area 3. As a result, the data in light gray is fetched onto cache one time, and the data in dark gray is fetched two times. It means that the cache miss rate is reduced drastically.

4 Pipelined Processing of Passes for EBCOT

Embedded block coding with optimal truncation (EBCOT) is the most complicated part in JPEG2000. The context of a sample coefficient is formed according to the significant state of the sample and its eight neighbors within 3x3 context window. Then the context data is processed by the arithmetic coder. Each bit-plane is encoded through three coding passes, called significant propagation pass (Pass 1), magnitude refinement pass (Pass 2) and clean up pass (Pass 3). In conventional EBCOT method, each pass is processed independently, although the processing of these three passes is very similar. Thus, if the redundancy between three passes, e.g. extraction of bit-plane data and context information from im-

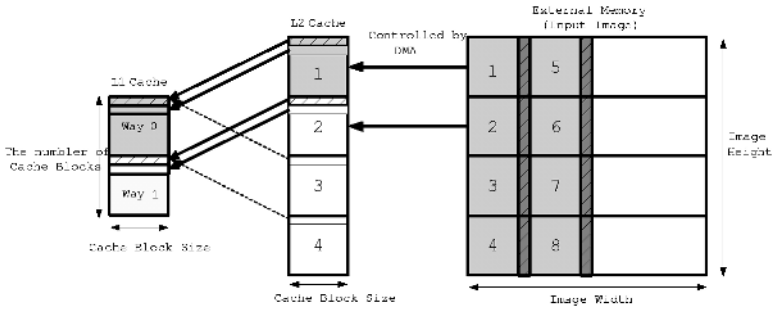


Fig. 4. Memory manipulation of proposed OBT.

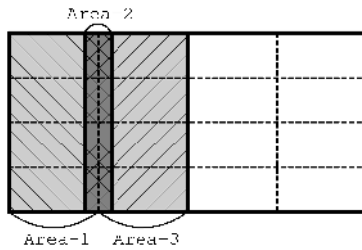


Fig. 5. Overlapped block configuration.

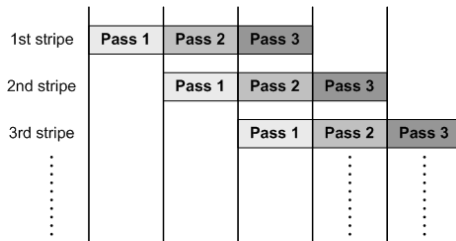


Fig. 6. Pipeline processing of three passes.

age data (16bits), is removed, it is possible to reduce remarkably the processing time of EBCOT.

Parallel processing can be utilized to remove the redundancy. However, due to the dependency of passes, it is difficult to process three passes of one stripe simultaneously. In detail, in order to process Pass 2 of current stripe, the context information of current and adjacent stripes, which is updated by processing Pass 1, is required. In case of Pass 3, the context information to be acquired by processing Pass 1 and Pass 2 is also needed.

So, we propose a pipelined processing of passes (PPP) scheme as an alternative method of parallel processing. The strategy is to process the three coding passes of the same bit-plane using pipeline architecture as shown in Fig. 6. First,

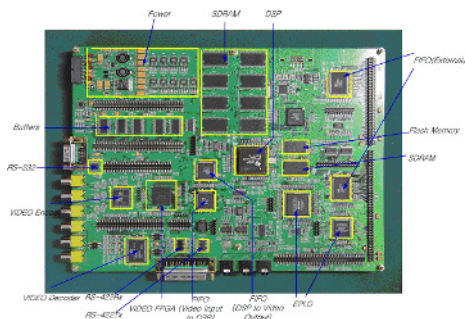


Fig. 7. The developed embedded Motion-JPEG2000 encoding system.

Table 1. Comparison of number of cache misses

Image size	Lifting direction	Number of cache misses	
		Conventional lifting	Overlapped Block Transferring
256x256	Horizontal	1024	1280
	Vertical	65,536	0
512x512	Horizontal	4096	4608
	Vertical	262,144	0

the bit-plane data of the first stripe is calculated for Pass 1. Second, the bit-plane data of the first and second stripes is calculated for Pass 2 and Pass 1 respectively. Third, the bit-plane data of the first, second and third stripes is processed for Pass 3, Pass 2 and Pass 1 respectively. Third step is iterated until the coding block is completely processed. As a consequence, all the three passes are processed in one scan.

5 Experimental Results

The proposed OBT-based lifting scheme and the PPP method for EBCOT are demonstrated in this section. To prove the effectiveness of the proposed method, simulations were conducted using on a TMS320C6416 (600Mhz, 4800MIPS). Fig. 7 shows the implemented Motion-JPEG2000 encoding system.

Table 1 shows the number of cache misses produced by the proposed OBT method. In the horizontal filtering, the OBT method produces more cache misses than the conventional method, because the data of overlapped area are fetched twice. However, in the vertical filtering, the OBT method completely removes the cache misses. Consequently, the OBT method reduces the cache-miss rate by 98%.

Table 2 shows the processing time of the DWT using the two existing method, which are Meerwald’s method (Row extension, Aggregation) [7] and Chatterjee’s method (Strip-mining, Data layout) [8], and the proposed OBT method. For all

Table 2. Comparison of processing time of wavelet lifting scheme

Different method		Execution time of DWT			
		Horizontal(ms)	Vertical(ms)	Total(ms)	Speed-up
Image size: 256x256					
Original wavelet-lifting		2.65	111.63	120.28	1
Meerwald's method	Row extension	2.85	24.66	27.51	4.38
	Aggregation	2.95	14.14	17.09	7.04
	Combination	2.88	10.88	13.76	8.74
Chatterjee's method	Strip-mining	2.71	32.27	33.98	3.54
	Data layout	2.87	41.12	43.99	2.76
	Combination	2.77	20.26	23.03	5.22
Overlapped block transferring		3.81	3.22	7.03	17.18

Table 3. Comparison of consuming time of EBCOT between conventional method and pipelined processing of passes.

Image		Lena	Baboon	Peppers
Conventional method (ms)	Pass 1	297.8	277.9	269.7
	Pass 2	140.3	156.8	157.2
	Pass 3	522.8	531.7	533.9
PPP method (ms)	Pass 1	298.6	281.6	272.8
	Pass 2	88.5	96.3	92.6
	Pass 3	357.4	369.5	378.3
Improvement		23%	25%	24%

image sizes, there is no improvement in the horizontal filtering. But all three methods are effective in vertical filtering. Row extension, aggregation and the combination of both methods reduce the processing time by 78, 88 and 90%, respectively, in the vertical filtering. Strip mining, recursive data layout and the combination of both methods reduce the processing time by 73, 66 and 82%, respectively, in the vertical filtering. Our method reduces the processing time by 98% in the vertical filtering. Note that the speed of horizontal filtering is almost identical to that of vertical filtering. It means that the proposed method eliminates most cache misses in vertical filtering, as we expected.

Table 3 shows the performance improvement by using the proposed PPP algorithm for EBCOT. As shown in Table 3, for Pass 1, the proposed method does not affect the execution time because there is no difference between the proposed method and the conventional method. However, for Pass 2 and 3, the proposed method reduces the calculation time up to 41% (Pass 2) and 32% (Pass 3). This result indicates that the proposed method significantly reduces the processing time for scanning and masking in case of Pass 2 and 3 by reusing the parameter and data used in Pass 1. Generally, the computation complexity of the whole EBCOT can be reduced by 24% as compared with the conventional architecture.

6 Conclusions

In this paper, we have presented a real-time embedded Motion-JPEG 2000 encoding system using a fixed-point DSP chip. To improve the performance of the system, we have proposed OBT-based lifting scheme to increase the cache hit rate. The OBT-based lifting scheme is over five times faster than the line-based lifting scheme. In addition, we showed that the proposed PPP algorithm can significantly reduce the execution time of EBCOT. Consequently, the Motion-JPEG2000 implementation on a DSP meets common requirement of real-time video coding [30 frames/s (fps)] and is proven to be a practical and efficient DSP solution.

References

- [1] Rabbani, M., Joshi, R.: An overview of the JPEG2000 still image compression standard. *Signal Processing and Image Communication*. 17 (2002) 3-48
- [2] Taubman, D., Marellin, M.W.: *JPEG2000: Image compression fundamentals, standards and practice*. Kluwer Academic Publishers. (2002)
- [3] Information Technology - JPEG2000 Image coding system:Part 1. ISO/IEC International Standard. 15444-1 (2000)
- [4] Yu, W., Qiu, R., Fritts, J.: Advantages of motion-JPEG2000 in video processing. in *Proceedings of the SPIE, Visual Communications and Image Processing*. 4671 (2002) 635-645
- [5] Daubechies, I., Sweldens, W.: Factoring wavelet transforms into lifting scheme. *The J. of Fourier Analsys and Applications*. 4 (1998) 247-269
- [6] Taubman, D.: High performance scalable image compressin with EBCOT. *IEEE Trans. Image Processing* 9 (2000) 1158-1170
- [7] Meerwald, P., Norecn, R., Uhl, A.: Cache issues with JPEG2000 wavelet lifting, *Proc. SPIE, Electron. Imaging, Vis. Commun. Image Process*. 4671 (2002) 626-634
- [8] Chatterjee, S., Brooks, C.D.: Cache-efficient wavelet lifting in JPEG2000. *IEEE Int. Conf. on Multimedia and Expo*. 1 (2002) 797-800

Low-Power Video Decoding for Mobile Multimedia Applications*

Seongsoo Lee and Min-Cheol Hong

School of Electronics Engineering, Soongsil University, Seoul 156-743 Korea
sslee@ssu.ac.kr

Abstract. This paper proposes a novel low-power video decoding scheme. In the encoded video bitstream, there are quite a large number of non-coded blocks. When the number of the non-coded blocks in a frame is known at the start of frame decoding, the workload of the video decoding can be estimated. Consequently, the supply voltage of VLSI circuits can be lowered, and the power consumption can be reduced. In the proposed scheme, the encoder counts the number of non-coded blocks and stores this information in the frame header of the bitstream. Simulation results show that the proposed scheme reduces the power consumption to about 1/20.

1 Introduction

Over the recent several years, mobile multimedia applications such as wireless videophone, mobile web terminal, and mobile multimedia broadcasting, are getting more and more important in the era of information technologies. Especially, they have become one of the killer applications in ubiquitous computing, which will impact our daily life in the near future.

In the hardware implementation of mobile multimedia terminal, power consumption is one of the primary concerns, since battery operation time is the key to commercial success [1]. Especially, it is essential to reduce the power consumption of video signal processing in mobile multimedia broadcasting applications such as digital multimedia broadcasting (DMB), since state-of-the-art batteries lasts less than one hour in mobile multimedia broadcasting terminals.

Most mobile multimedia broadcasting applications include video decoding, audio decoding, and other system functions. Fig. 1 (a) shows the power consumption of a typical mobile multimedia broadcasting, where video decoding is dominant in power consumption. In the video decoding, macroblock-level decoding including macroblock header decoding, inverse quantization (IQ) and inverse discrete cosine transform (IDCT) dominates total power consumption, as shown in Fig. 1(b). Consequently, low-power macroblock-level video decoding is essential in mobile multimedia broadcasting terminals.

Recently, dynamic voltage scaling (DVS) [2] efficiently reduces the power consumption of VLSI circuits, where supply voltage of VLSI circuits is dynamically

* This work was supported by the Soongsil University Research Fund.

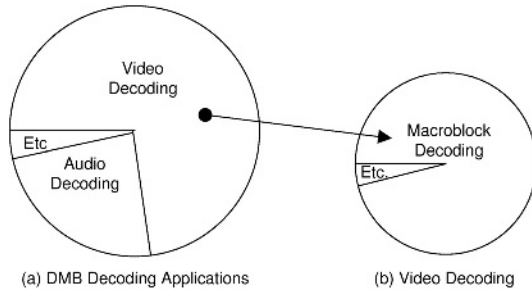


Fig. 1. Power consumption in DMB decoding

lowered to the lowest possible extent so as not to violate real-time execution of given applications. It is known as one the most efficient methods to reduce the power consumption of multimedia signal processing. However, algorithmic support of multimedia signal processing is still needed for efficient power reduction, since there are many problems when dynamic voltage scaling is directly applied to multimedia signal processing.

This paper proposes new low-power video decoding called *Zero-Block Skipping Macroblock Decoding (ZBSMD)*, where information of uncoded blocks and macroblocks is stored in the video bitstream. This information is exploited by the video decoder, and the supply voltage of VLSI circuits is lowered to the lowest possible extent for the frame-level real-time video decoding.

2 Dynamic Voltage Scaling

Power consumption of VLSI circuits is given by $P = \alpha C_L V^2 f$, where α is the switching activity, C_L is the load capacitance, V is the supply voltage, and f is the clock frequency [2]. In general, it is the most effective to lower the supply voltage to reduce the power consumption, since the power consumption is proportional to the square of the supply voltage. However, the circuit delay is given by $T \propto \frac{V}{(V - V_{TH})^\alpha}$, where V is the supply voltage, V_{TH} is the threshold voltage, and α is the velocity saturation index (~ 1.3) [3]. Consequently, as shown in Fig. 2 (a), circuit delay increases and its corresponding system operating frequency decreases when the supply voltage is lowered.

When the required workload of given applications is lower then the maximum system throughput, the system operating frequency can be lowered to the extent such that the system finishes given applications just at the deadline. The supply voltage can be lowered to that extent, and the power consumption reduces dramatically. This is the key idea of the dynamic voltage scaling (DVS) [2].

Fig. 3 illustrates the dynamic voltage scaling in detail. When the operating frequency is 50MHz and the workload of given task is 5×10^8 cycle, it finishes computing at $t = 10$ and idles until $t = 25$. Assuming that the power

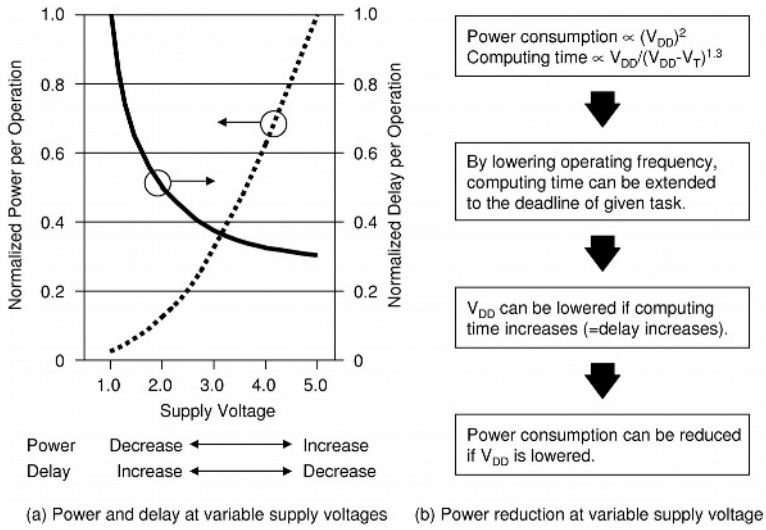


Fig. 2. Power reduction using variable supply voltage

per cycle = 1nJ/cycle and supply voltage = 5.0V, total power consumption = 0.5J. However, when the supply voltage and the clock frequency are lowered to 4.0V@40MHz and 1.5V@15MHz, the power per cycle = 0.64nJ/cycle@4.0V and 0.09nJ/cycle@1.5V, and the total power consumption = 0.155J. Note that it finishes computing at $t = 25$ and never idles since the supply voltage is lowered until it finishes computing just at its deadline. Similarly, when the supply voltage and the operating frequency are lowered to 2.0V@20MHz, total power consumption = 0.08J, achieving minimum power consumption.

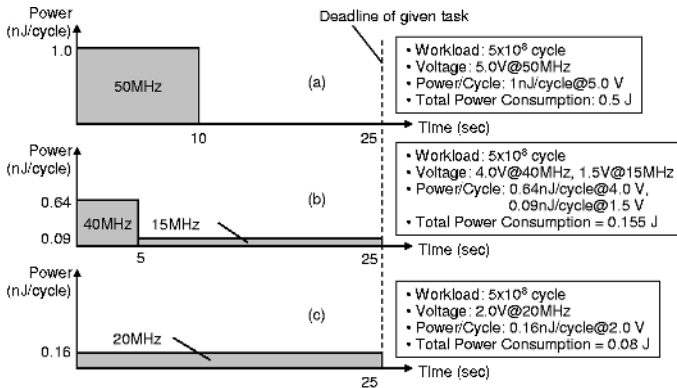


Fig. 3. Dynamic voltage scaling

3 Zero-Block Skipping Macroblock Decoding

In MPEG-4 video coding [4], some DCT coefficients are truncated to zero in the quantization process. If all 64 DCT coefficients of 8×8 pixels are truncated to zero, that block is non-coded, and the decoder doesn't need to perform inverse quantization (IQ) nor inverse DCT (IDCT) for that block. If all 6 blocks are non-coded, only motion vector is coded and transmitted, and the decoder doesn't need to perform IQ nor IDCT for whole macroblock. However, this non-coded block information is stored in *cbpy* and *mcbpc* fields of encoded bitstream. These fields locate in the head of each macroblock, and the decoder knows whether to skip the blocks only when it starts to decode each macroblock.

To apply dynamic voltage scaling, the decoder decodes *cbpy* and *mcbpc* fields first, and finds out the number of non-coded blocks in the macroblock, and determines proper supply voltage for the macroblock. Assume that the decoder processes CIF sequences (352×288 pixels, 30 frames/s). There are 396 macroblocks in a frame, and the time interval for each macroblock is $84.2 \mu\text{s}$. Since the decoder doesn't know the non-coded block information of next macroblocks, the decoder should finish each macroblock within $84.2 \mu\text{s}$ deadline. Assume that the transition delay to change supply voltage is zero, and current macroblock has only 3 coded blocks (block 0, 3, and 5) as shown in Fig. 4 (a). When dynamic voltage scaling is not applied, the decoder runs and stops as in Fig. 4 (b), and the supply voltage and its corresponding power consumption are determined as in Fig. 4 (c) and (d). Note that the motion compensation (MC) is always performed even if the block is non-coded. When dynamic voltage scaling is applied, the workload is $\frac{1}{2}$ compared when all blocks are coded, and the supply voltage can be lowered about $\frac{1}{2}$ of maximum supply voltage. In this case, the supply voltage and its corresponding power consumption is determined as in Fig. 4 (e) and (f). The power is proportional to the clock frequency and the square of the supply voltage. The power is reduced to $\frac{1}{8}$, since the supply voltage is reduced to $\frac{1}{2}$ and the clock frequency is also reduced to $\frac{1}{2}$. Consequently, the power consumption is reduced to $\frac{1}{4}$.

However, this dynamic voltage scaling scheme has a critical problem. According to the characteristics of VLSI circuits, it takes quite a long time to change supply voltage of VLSI circuits. Even state-of-the-art DC-DC converter circuits take about $100 \mu\text{s}$ to change supply voltage. Since the assigned time for each macroblock is only $84.2 \mu\text{s}$, and it seems almost impossible to change supply voltage several times during macroblock processing.

We solved the problem by storing non-coded block information in the frame header. In MPEG-4 video coding, user data can be stored in the bitstream using *pei* and *psupp* fields. These fields are usually ignored and they make no effects in conventional decoders. In the proposed *Zero-Block Skipping Macroblock Decoding* (ZBSMD), the encoder counts the total number of non-coded blocks in a frame, and this information is stored in the *pei* and *psupp* fields in the frame header. In the decoder, this information is decoded at the start of frame decoding, and the decoder knows the number of non-coded blocks in a frame before frame processing. In this case, the decoder knows the exact workload of

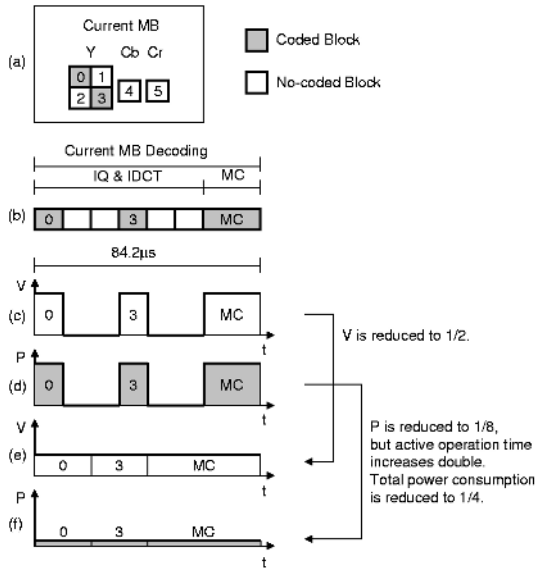


Fig. 4. Power reduction of dynamic voltage scaling in video decoding when transition time to change supply voltage is zero

frame processing, and the decoder determines proper system operating frequency and its corresponding supply voltage.

This supply voltage changes only once at the start of frame processing, and doesn't change during it. The transition time to change supply voltage ($\sim 100\mu s$) is negligible compared to frame processing time ($= 33333.3 \mu s$), and the supply voltage can be lowered to reduce power consumption. This scheme can be easily adapted to H.264 [5] and other video coding standards.

Fig. 5 illustrates the proposed scheme in detail. As shown in Fig. 5 (a), the encoder encodes the reference frame. At the same time, the encoder counts the number of non-coded blocks. After the encoder finishes encoding the reference frame, the encoder inserts the number of non-encoded blocks into the encoded bitstream. The additional data added to the bitstream is 18 bits per frame (2 *pei* bits and 16 *psupp* bits), since the number of blocks in a frame is 2376 in CIF format and 9504 in CCIR601 format. This increases 0.0054 % of the encoded bits when the bitrate is 1 Mbits/s and the frame rate is 30 frames/s, which is negligible.

As shown in Fig. 5 (b), the decoder decodes the frame header first, and then reads the number of non-coded blocks. After that, the decoder calculates the proper supply voltage to finish the decoding of macroblock data just at given deadline, and the lowered supply voltage is applied to the decoder to reduce power consumption. When the decoder doesn't support dynamic voltage scaling,

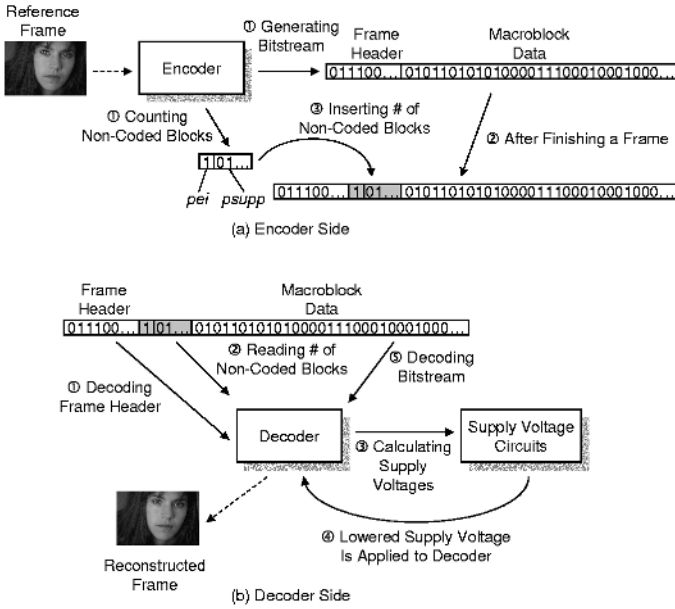


Fig. 5. The proposed zero-block skipping video decoding scheme

the number of non-coded blocks in *pei* and *psupp* fields are just ignored, and the decoder performs conventional video decoding.

The proper supply voltage for the proposed scheme is calculated as follows. The required system operating times to decode a block and to perform motion compensation of a macroblock are t_B and t_{MC} , respectively. N is the total number of blocks in a frame. N_{NC} is the number of non-coded blocks in a frame, and it is stored in the frame header. f_{MAX} and V_{MAX} are the maximum system operating frequency and maximum supply voltage to decode a frame when there are no non-coded blocks, respectively. f_{SYS} and V_{SYS} are the required system operating frequency and the required supply voltage to decode a frame when there are N_{NC} non-coded block, respectively. Assuming that the decoding time for frame header is negligible, f_{SYS} is calculated as Eqn. (1). Note that f_{SYS} decreases as N_{NC} increases, since the required computation decreases as the number of non-coded blocks increases. When f_{SYS} decreases, the required supply voltage is lowered and the power consumption is reduced. Note that Eqn. (1) is calculated only once per each frame, and the computational overhead is negligible. After f_{SYS} is calculated, V_{SYS} is calculated as Eqn. (2), since $T = \frac{1}{f} \propto \frac{V}{(V - V_{TH})^\alpha}$, where V is the supply voltage, V_{TH} is the threshold voltage, and α is the velocity saturation index (~ 1.3) [3]. Since f_{MAX} , V_{MAX} , V_{TH} , and α are constant for given VLSI circuits, V_{SYS} can be pre-calculated and tabulated as a lookup table of f_{SYS} . Note that the computational overhead for Eqn. (2) is also negligible, since it is performed by table look-up.

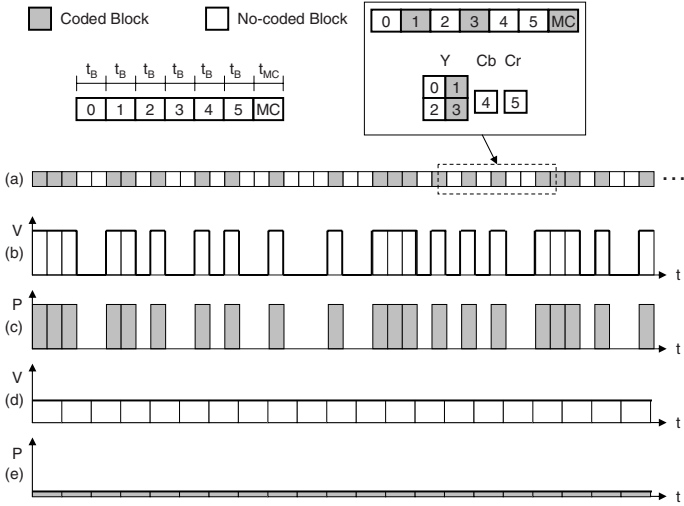


Fig. 6. Power reduction of the proposed zero-block skipping video decoding scheme

Table 1. The normalized number of non-coded blocks, the normalized supply voltage, and the normalized power consumption.

Parameters	Claire	Miss America
Normalized number of non-coded blocks	0.66	0.65
Normalized supply voltage	0.24	0.25
Normalized power consumption	0.054	0.057

$$f_{SYS} = f_{MAX} \times \frac{(N - N_{NC}) \times t_b + \frac{N}{6} \times t_{MC}}{N \times t_b + \frac{N}{6} \times t_{MC}} \tag{1}$$

$$f_{MAX} \times \frac{V_{SYS}}{(V_{SYS} - V_{TH})^\alpha} = f_{SYS} \times \frac{V_{MAX}}{(V_{MAX} - V_{TH})^\alpha} \tag{2}$$

Fig. 6 illustrates the power reduction of the proposed scheme when $t_B = t_{MC}$ and $N_{NC} = \frac{5N}{12}$. When dynamic voltage scaling is not applied, the decoder runs and stops as in Fig. 6 (a), and the supply voltage and its corresponding power consumption are determined as in Fig. 6 (b) and (c). When dynamic voltage scaling is applied, $f_{SYS} = \frac{1}{2}f_{MAX}$ and $V_{SYS} \sim \frac{1}{2}V_{MAX}$, and the supply voltage and its corresponding power consumption is determined as in Fig. 6 (d) and (e). The power is proportional to the clock frequency and the square of the supply voltage. The power is reduced to $\frac{1}{8}$, since the supply voltage is reduced to $\frac{1}{2}$ and the clock frequency is also reduced to $\frac{1}{2}$. Consequently, the power consumption is reduced to $\frac{1}{4}$.

In this paper, MPEG-4 SP@L2 video coding was tested with two CIF sequences (352×288 pixels, 30 frames/s), i.e “Claire” and “Miss America”. The

supply voltage V_{DD} , the threshold voltage V_{TH} , and the velocity saturation index α are 2.5V, 0.5V, and 1.3, respectively. Transition delay to change supply voltage is 100 μ s.

Table 1 shows the normalized number of non-coded blocks $\frac{N_{NC}}{N}$, the normalized supply voltage $\frac{V_{SYS}}{V_{MAX}}$, and the normalized power consumption $\frac{P_{SYS}}{P_{MAX}}$. As shown in Table 1, non-coded blocks are about 65% of total blocks in a frame. The supply voltage is lowered to about 25%, and the power consumption is reduced to about 5%.

4 Conclusion

This paper presents a novel low-power video decoding scheme. In the video bitstream, some blocks are non-coded, and the decoder can slow down the system operating frequency. By exploiting dynamic voltage scaling, the supply voltage can be lowered and the power consumption can be reduced. However, the non-coded block information is known only at the start of macroblock decoding, which adjusts the supply voltage so often. Usually the transition delay to change supply voltage is so large that the supply voltage cannot be adjusted at macroblock level. In the proposed scheme, the encoder counts the number of the non-coded blocks in a frame, and inserts this information into the bitstream. Using this information, the decoder can calculate the proper supply voltage. The supply voltage is lowered to a single voltage for entire frame processing, and the power consumption is reduced significantly. The proposed scheme needs neither complex algorithm nor computational overhead. Furthermore, the proposed scheme can be easily adapted to most existing video coding standards. When applied to the real-time multimedia applications, it reduces the power consumption to about 1/20 without degrading performance.

References

1. Rabaey, J.: Low-Power Silicon Architectures for Wireless Communication, Asia and South Pacific Design Automation Conference (2000) 379–380.
2. Chandrakasan, A., Brodersen, R.: Low Power Digital CMOS Design, Kluwer Academic Publishers (1995).
3. Sakurai, T., Newton, A.: Alpha-Power Law MOSFET Model and Its Application to CMOS Inverter Delay and Other Formulas, IEEE Journal of Solid State Circuits **25** (1990) 584-594.
4. ISO/IEC JTC1/SC29/WG11 14496-2: Coding of Audiovisual Object: Visual (1998).
5. ITU-T Rec. H.264: Advanced Video Coding (2002).

Adaptive Macro Motion Vector Quantization

Luis A. da Silva Cruz*

Dep. of Electrical and Computer Engineering
Polo 2 University of Coimbra
3030-290 Coimbra, Portugal
lcruz@deec.uc.pt

Abstract. Block based motion estimation can be interpreted as a form of vector quantization (VQ) based on a codebook of motion vectors. As is usually the case in VQ the codebook can be precomputed using the Lloyd-Buzzo-Gray algorithm and a set of training data or generated on-the-fly in an adaptive fashion. In this work we present an adaptive motion estimation method based on vector quantization of blocks of motion vectors with a dynamically updated codebook which reduces the complexity of the motion estimation process and the motion side information.

Keywords: Motion estimation, vector quantization, video coding.

1 Introduction

To our knowledge the first recorded association of block-matching motion estimation and vector quantization appeared in [1]. The researcher applied the well studied vector quantization (VQ) methods [2] to design a codebook of motion vectors which was used to quantize the motion vectors obtained by full search block-matching (FSBM) motion estimation. In [3] Lee and Woods improved the idea by changing the distortion measure used in the clustering process from a measure based on the Euclidean distance between the motion vector to be coded and the candidate codevector, to a measure based on the distortion induced by the motion vectors and measured in the image domain, i.e. the matching distortion. Later in [4] and [5] several modifications were introduced making the method adaptive. In [4] adaptation was introduced in the form of a procedure which updated the codebook by replacing the least frequently used codevectors (*LFU* strategy) or the codevectors which had been unused for the longest time (Least Recently Used strategy or *LRU*) by better vectors. This method does not require transmission of codebook adaptation side information since the information necessary to perform the codebook adjustment is available at both the coder and decoder. We called this method *Backward Adaptive Motion VQ* (BAMVQ).

In [5] a variation of the previous method was presented in which the motion vector codebook is periodically partially retrained, with the worst codevectors

* The author wishes to thank FCT-Fundação para a Ciência e a Tecnologia and IT-Instituto de Telecomunicações Pólo de Coimbra for the support provided.

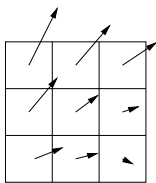


Fig. 1. Macro motion vector, 3x3 geometry

being updated through an Loyd-Buzzo-Gray (LBG) iterative process based on data from the current and previous frames. Given that the original current video frame is available only at the coder, this method requires transmission of codebook update information. This method was called *Forward Adaptive Motion VQ* or FAMVQ.

In all of the previously mentioned methods, the codevectors consisted of a single motion vector and here we present an extension to codevectors consisting of several motion vectors.

1.1 Motion Vector Blocking

Henceforth *macro motion vector* will denote a set of motion vectors whose corresponding image blocks are spatially arranged as exemplified in figure 1 for the case of a 3 by 3 macro vector. By *MMVQ codebook* we mean a set of such *macro motion vectors*. In the work to be presented the underlying image blocks are groups of 8 by 8 pixels so that a 3 by 3 macro vector corresponds to a 24 by 24 block of pixels (macroblock).

More formally we will call the macro motion vector codebook, \mathcal{C} , with M macro motion vectors $mmv_i, i = 0, \dots, M - 1$, where mmv_i is a block of motion vectors,

$$mmv_i = \{mv_{i,0}, mv_{i,1}, \dots, mv_{i,S-1}\}$$

and S is the number of motion vectors per macro motion vector. An image frame will be decomposed into MB non-overlapping macroblocks, $mb_j, j = 0, \dots, MB - 1$. Each macroblock is an assembly of S blocks,

$$mb_j = \{b_{j,0}, \dots, b_{j,S-1}\}.$$

The distortion incurred when we match macroblock mb_j in the current frame against the same position macroblock in the previous reconstructed frame after motion compensation using macro motion vector mmv_i is represented by $md(mb_j, mmv_i)$ and is defined to be

$$md(mb_j, mmv_i) = \sum_{v=0}^{S-1} d(b_{j,v}, mv_{i,v})$$

where $d(b, mv)$ is the mean of absolute differences (MAD) between block b and the block from the previous image frame with displacement relative to block b given by mv .

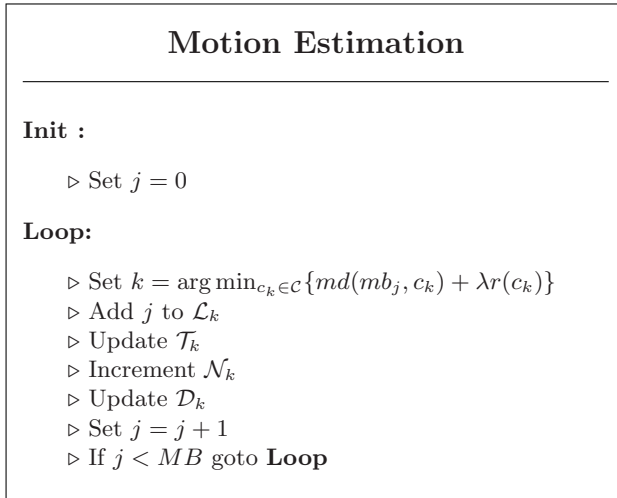


Fig. 2. Motion Estimation - MMVQ

1.2 Backward Macro Motion VQ

The motion estimation algorithm presented here can be decomposed into two stages, *motion estimation* and *codebook update*. The motion estimation operations are summarized by the pseudocode in figure 2. The estimation of the motion of image macroblock j amounts to a search over the the codebook for the macrovector mmv_i which yields the lowest value for the weighted sum of the matching distortion $md(mb_j, mmv_i)$ and a measure of the expected codevector index entropy $r(mm v_i)$ estimated from the histogram of past codevector use. \mathcal{T}_k is used to keep track of codevector's k last time of use, in \mathcal{N}_k we keep the number of macroblocks clustered to codevector k , in \mathcal{D}_k the respective cumulative distortion and in \mathcal{L}_k the clustering information in the form of a list of macroblocks clustered to codevector k . After a given number of macroblocks has had its motion estimated (in this work this occurs at video frame intervals), we enter the codebook update stage, which uses the clustering information from previous motion estimation steps to alter the codebook composition. The codebook update procedure is based on the assumption that codevectors that haven't been used for a long time (*LRU* strategy) or that haven't been used as often as the others (*LFU* strategy), should be replaced. In BAMMVQ these less useful codevectors are replaced by variations of the most useful codevectors, where the variation is introduced by the addition of a pseudo-random perturbation codevector to the source codevector. The source macrovector is chosen from the macrovectors with highest use (highest cluster occupation). This criterion assumes that macrovectors with high use are better macrovectors. Other criteria such as average cluster distortion were tried but the results obtained were worse. Experimentation showed that the new codevector generation scheme just

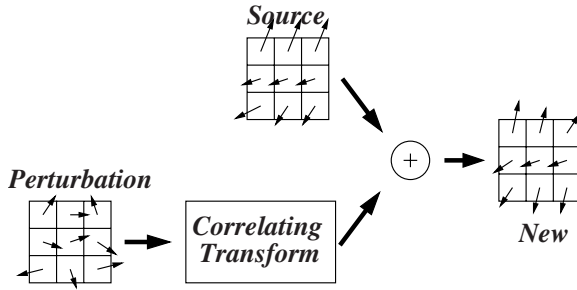


Fig. 3. Random macro vector generator

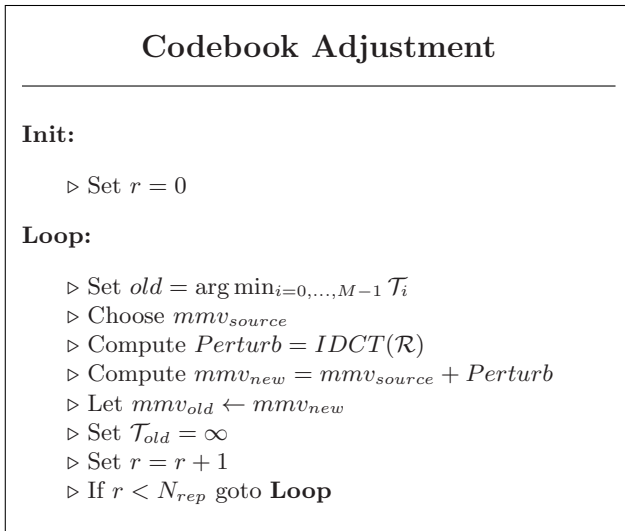


Fig. 4. Codebook update - BAMMVQ

described did not work well, and the addition of the pseudo-random perturbation had to be preceded by a *correlating transform* as illustrated in figure 3. In this work we used the Inverse Discrete Cosine Transform (IDCT) as the correlating transform. The complete codebook adaptation procedure is specified in pseudocode in figure 4, where N_{rep} represents the number of codevectors to be updated, \mathcal{R} is a macrovector of random values obtained as described before and the criterion used to select the codevector to be replaced is *LRU*.

1.3 Forward Macro Motion VQ

The forward adaptive macro motion VQ (FAMMVQ) algorithm is similar to BAMMVQ described in the previous section. The codebook is used exactly in

the same manner during motion estimation and it is only the adaptation procedure that differs. Similarly to BAMMVQ, in FAMMVQ after a series of motion estimation steps, we replace some macrovectors which were tagged as needing improvement by better ones. For each macrovector that we want to replace, we select a source macro motion vector, derive a new macro motion vector from it by addition of a perturbation macrovector and then using the macroblocks clustered to the source macrovector execute a partial retraining of the source and new macrovectors using the *k-means* algorithm [2]. To compute the new centroid for each of the two clusters (associated with the source and new macrovectors) each component vector of the source and new macrovector is improved by exhaustive search over a small window centered on the initial value of the component vector. This method of centroid determination, although guaranteeing that the true centroid is found, translates into high computational costs. To reduce the number of operations needed in this retraining step we perform the extended search (over *Search Region 1*) only for the centroids of the corner vectors of the macrovectors and then use those values to interpolate initial vectors for the remaining macrovector positions. The vectors with these initial values are then refined by a search over a much smaller region (*Search Region 2*). We verified experimentally that the performance decrease incurred by this shortcut is negligible. The data used in the retraining of source macrovector i and the new macrovector are all the macroblocks which were clustered to macro motion vector i whose identities were stored (during the motion estimation phase) in the list \mathcal{L}_i . Once we have obtained the desired improved macrovectors, we have to transmit their values to the decoder, so that he can update his copy of the codebook. We do so by PCM encoding the top-left vector of each macrovector and send all other vectors of the same macrovectors encoded differentially in left-to-right top-to-bottom order. This encoding is applied to the horizontal and vertical components of each vector. Since the amount of adaptation information was not very large we did not use any kind of entropy coding.

2 Simulation Setup and Results

2.1 Coder Framework

In order to quantify the performance of the algorithm proposed, we built a full color motion compensated predictive coder which uses our algorithm in the motion estimator, and a SPIHT kernel to code the luminance and chrominance channels. Since the target video data is in interlaced format (ITU-R 601), we use a motion estimator which first tries to estimate the motion of the current frame with respect to the previous reconstructed frame, and if the motion compensated residue energy is above a given threshold (obtained empirically) the motion is reestimated, this time at field level. A bit is sent to indicate which mode (*field* or *frame*) was used for the current frame. The codevector indices coming out of the motion estimator are entropy coded conditioned on the previous index using adaptive arithmetic coding. The motion estimation and compensation is

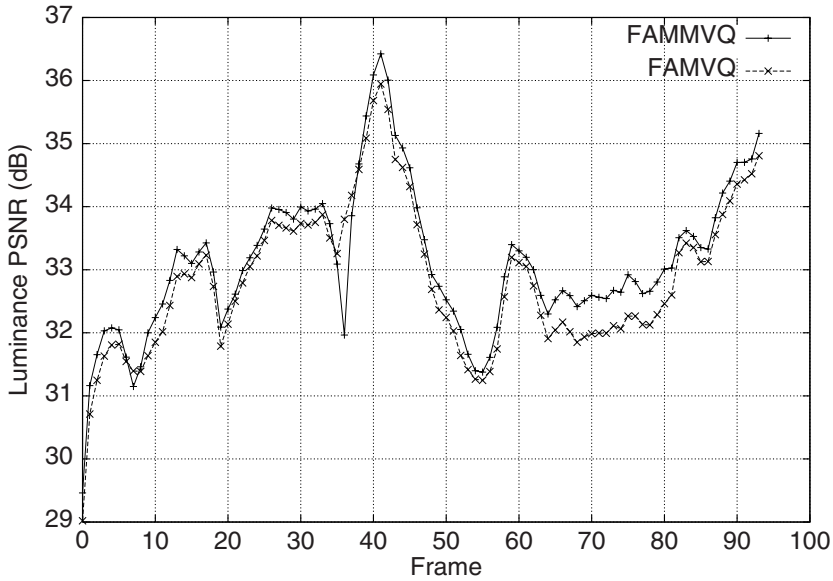


Fig. 5. FAMMVQ versus FAMVQ - *Mobile* - 0.75 bpp

performed at half-pel accuracy, and so all the vectors in the macrovectors support this same precision. In the case of FAMMVQ, the codebook adaptation information is sent multiplexed with the compressed video information.

2.2 Results

Concerning the backward schemes BAMVQ and BAMMVQ, the experiments we performed showed that BAMMVQ was consistently worse than BAMVQ. Possible justifications for this fact are discussed later. As for FAMMVQ in figure 5 we show results comparing FAMMVQ to FAMVQ when coding a section of the *Mobile* sequence, also for a codebook containing 128 single motion vectors for FAMVQ and 128 4x4 macro motion vectors for FAMMVQ and an aggregated bitrate of 0.75 bpp where we can observe a slight increase in performance of FAMMVQ when compared to FAMVQ. Similar results were observed for other sequences and rates.

In table 1 we present the mean values of the motion bitrate obtained when coding 60 frames of sequence *Table Tennis* for all the competing motion estimation methods under study (FSBM, BAMVQ, FAMVQ, BAMMVQ and FAMMVQ). The values presented for FAMVQ and FAMMVQ include the bits used to code the codebook adaptation information. All rates are expressed in bits per pel (bpp).

At the heart of the FAMMVQ adaptation process is a partial retraining algorithm which operates iteratively until convergence is attained. Since these iterations are computationally expensive, it is important to know how fast the

Table 1. Motion Bitrate (in bpp) for 60 frames of sequence *Table Tennis*

Total Rate (bpp)	0.50	0.75	1.00	1.50
FSBM	0.0234	0.0233	0.0227	0.0225
BAMVQ	0.0182	0.0182	0.0181	0.0178
FAMVQ	0.0178	0.0178	0.0178	0.0177
BAMMVQ	0.0033	0.0031	0.0030	0.0028
FAMMVQ	0.0035	0.0033	0.0032	0.0029

process converges. Empirical results shows that about 80% of the retraining steps converge within 2 iterations of the LBG cycle. Regarding image quality, subjective evaluation of reconstructed video seems to evidence some advantage towards FAMVVQ when compared to FAMVQ, as the former results appear more stable and less noisy, specially at lower (total) bit rates.

3 Conclusions

The results obtained show that contrary to what was expected, extending Backward Adaptive MVQ (BAMVQ) to multidimensional (macro) vectors (BAMMVQ) did not bring any improvement in performance. We believe that this is due to the simplicity of the adaptation method, which is based on the addition of a random vector to a *good* vector, as described before.

In the single motion vector codevector case, this method worked reasonably well, since the odds of getting a resulting good vector using the method described are high. In the case of multidimensional codevectors, the result codevector universe is much larger (due to the higher dimensionality) and so it is not as likely that adding a (multidimensional) random perturbation to a good macrovector will result in an equally good macrovector. In the case of forward adaptation we observed an improvement in the performance when going from single motion vector codevectors to multi motion vector codevectors, as can be observed for instance in PSNR plot in figure 5. This is most certainly due to the effect of the partial retraining included in FAMMVQ which improves the new macrovectors generated (as in BAMMVQ) by the addition of a random perturbation to a good macrovector, therefore eliminating the dimensionality problem described before for the case of the backward adaptive algorithm (BAMMVQ).

The methods presented in this work can be viewed as fast block based motion estimation procedures considering the small size of the codebook when compared to the number of search points in full search block matching motion estimation and also as a motion field adaptive vector quantization method wich achieves a motion field representation with a reduced number of bits when compared to full search block match. Concerning the complexity of FAMMVQ (the most computationally demanding of the two) we have a clear advantage over FSBM. Indeed for ITU-R 601 video sequences, FSBM with block size 8 and search region ± 11 at full pel resolution plus half pel refinement we need about 3.5 million block

compares whereas FAMMVQ with block size 8 macroblock geometry 4 by 4 and codebook with 128 macro motion vectors requires about 830 thousand block compares. The cost of codebook adaptation with 13 codevectors updated at each adaptation event is about 1.8 million block compares so that the total cost of motion estimation and codebook adaptation in FAMMVQ is smaller than the cost of FSBM motion estimation. Even though the results presented are somewhat modest, we think the technique introduced here (AMMVQ) deserves further investigation which might lead to improved performance. Work is ongoing on the adaptation of the method presented in this document to the H.264/AVC standard. We believe that use of a codebook based approach such as the one presented here can significantly reduce the cost of searching the very large motion vector space characteristic of the multi reference motion compensated prediction employed in H.264/AVC. Faster codebook adaptation procedures are also being investigated as well as an extension of the method to use more general shapes of motion vector groups. Concurrently a multiresolution version of the method (macro motion VQ) is being developed which uses a separate macro motion vector codebook for each resolution level.

References

1. Michael Gilge, "Motion estimation by scene adaptive block matching (SABM) and illumination correction," in *Proceedings of the SPIE Conference on Image Processing Algorithms and Techniques*, 1990, vol. 1244, pp. 355–366.
2. Allen Gersho and Robert M. Gray, *Vector quantization and signal compression*, Kluwer Academic Publishers, 1992.
3. Yoon Yung Lee and John W. Woods, "Motion vector quantization for video coding," *IEEE Transactions on Image Processing*, vol. 4, pp. 378–382, March 1995.
4. Luis A. da Silva Cruz and John W. Woods, "Backward adaptive motion vector VQ for video coding," in *Proceedings of the Picture Coding Symposium*, April 1999, pp. 25–28.
5. Luis A. da Silva Cruz and John W. Woods, "Adaptive motion vector VQ for video coding," in *Proceedings of the IEEE International Conference on Image Processing*, September 2000, pp. 867–870.

Automatic, Effective, and Efficient 3D Face Reconstruction from Arbitrary View Image

Changhu Wang^{1,3}, Shuicheng Yan², Hua Li², Hongjiang Zhang³, and Mingjing Li³

¹ Department of EEIS, University of Science and Technology of China
wch@ustc.edu

² LMAM, School of Mathematical Sciences, Peking University, P.R. China
{scyan, lihua}@math.pku.edu.cn

³ Microsoft Research Asia, Beijing, P.R. China
{hjzhang, mjli}@microsoft.com

Abstract. In this paper, we propose a fully automatic, effective and efficient framework for 3D face reconstruction based on a single face image in arbitrary view. First, a multi-view face alignment algorithm localizes the face feature points, and then EM algorithm is applied to derive the optimal 3D shape and position parameters. Moreover, the unit quaternion based pose representation is proposed for efficient 3D pose parameter optimization. Compared with other related works, this framework has the following advantages: 1) it is fully automatic, and only one single face image in arbitrary view is required; 2) EM algorithm and unit quaternion based pose representation are integrated for efficient shape and position parameters estimation; 3) the correspondence between 2D contour points and 3D model vertexes are dynamically determined by normal direction constraints, which facilitates the 3D reconstruction from arbitrary view image; 4) a weighted optimization strategy is applied for more robust parameter estimation. The experimental results show the effectiveness of our framework for 3D face reconstruction.

1 Introduction

Modeling 3D human faces has been a challenging problem in computer graphics and computer vision literatures in the last decades. Since the pioneering work of Parke [8][9], various techniques have been reported for modeling the geometry of faces [5][11]. The 2D-based methods do not consider the specific structure of human faces, thus result in the poor performance on profile face samples. In the work of Lam et al. [4], face samples with out-of-plane rotation are warped into frontal faces based on a cylinder face model, but it requires heavy manual labeling work. Shape from shading [13] has been explored to extract 3D face geometry information and generate virtual samples by rotating the generated 3D models. This algorithm requires that the face images are precisely aligned pixel-wise, which is difficult to be implemented in practice and even impossible for practical face recognition applications.

The two most popular works on 3D face modeling and analysis are the morphable 3D face model proposed by Vetter et al. [10] and the artificial 3D shape model by Zhang et al. [7]. The former one presented a 3D face reconstruction algorithm to recover the shape and texture parameters based on a face image in arbitrary view, and the latter developed a system to construct textured 3D face model from video sequence. Recently, Hu and Yan et al. [3] presented an automatic 2D-to-3D integrated face reconstruction method to recover the 3D face model based on a frontal face image and it is much faster. However, there are still some shortcomings in these works: 1) both Vetter and Zhang's works require manual initialization and the speed can not satisfy the requirement of a practical face recognition system; 2) Vetter's work needs a large number of samples for a representative texture model, and mostly the small number of texture samples will limit the generalization of the algorithm; 3) Hu and Yan's work assumed fixed pose parameters which limited its extension to side view images.

In this paper, we propose a fully automatic, effective and efficient framework for 3D face reconstruction based on a single face image in arbitrary view. It not only inherits the advantages of the above three works, but also successfully overcomes their shortcomings. First, a recently developed multi-view face alignment algorithm [6] is utilized to localize the feature points in a face; Second, the 3D face shape and pose parameters are estimated synchronously by an EM based algorithm, in which the correspondence between the contour points and their vertex indices in the 3D face models are dynamically determined; moreover, a unit quaternion based pose representation is proposed for efficient position parameter optimization; Finally, the complete 3D face model is obtained by mapping the input 2D image onto the 3D face shape surface with the mirror and smoothing operation.

The rest of this paper is organized as follows. The 2D-to-3D face reconstruction algorithm is described in detail in Section 2. Section 3 provides some experimental results. We draw the conclusions and discuss the future work in Section 4.

2 3D Reconstruction with Single Arbitrary View Image

In this section, we present our fully automatic framework for 3D face reconstruction. In [3], Hu and Yan et al. proposed an automatic algorithm for 3D face reconstruction; however, it can only handle frontal faces. In this framework, we utilize a newly developed multi-view face alignment algorithm [6] to locate the feature points in an arbitrary view face image; then, the 3D shape and position parameters are efficiently estimated with the EM algorithm in term of the unit quaternion [1][2] pose representation and the dynamical correspondence between the contour points and the vertexes on the 3D face model. Moreover, a weighted optimization strategy is applied for robust parameter estimation. This section consists of five parts: 1) the efficient multi-view 2D face alignment; 2) the morphable 3D face model; 3) problem formulation; 4) efficient parameter inference; and 5) robust parameter estimation with the dynamic correspondence strategy and weighted optimization.

2.1 Efficient Multi-view 2D Face Alignment

Automatic multi-view face alignment is still an open problem. In this work, we apply the recently proposed multi-view 2D alignment algorithm [6]. In [6], the texture is redefined as the unwrapped grey-level edge in the original image; then, a Bayesian network is designed to describe the intrinsic co-constraints between shape and texture; finally, the EM algorithm is utilized to infer the optimal parameters of the proposed Texture-Driven Shape Model. There are 83 feature points located, part of which are adaptively selected for 3D face reconstruction in different views.

2.2 Morphable 3D Face Model

Similar to Vetter's work [10], the geometry of a 3D face is represented as a shape vector $S = (x_1, y_1, z_1, x_2, \dots, y_L, z_L)' \in \mathcal{R}^{3L}$, which contains the x , y and z coordinates of the L vertices. We apply the probabilistic extension of traditional PCA [12] to model the shape variations based on 100 3D faces with about 8900 vertexes.

$$S = U \cdot s + \bar{S} + \varepsilon, \quad \varepsilon \sim N(0, \sigma_{3d}^2 I_{3L}), \quad \sigma_{3d}^2 = \sum_{i=l+1}^{3L} \lambda_i / 3L \quad (1)$$

where the columns of U are the most significant eigenvectors and l is the number of eigenvectors, \bar{S} is the average shape of samples and s is the shape parameter to be estimated. ε denotes the isotropic noise in the shape space and σ_{3d} is the standard deviation.

2.3 3D Reconstruction Problem Formulations

The input is the multi-view face alignment result as described in subsection 2.1, denoted as s_{2d} , and the object is to reconstruct the personalized 3D face model. Their relationship can be formulated as:

$$s_{2d} = PfRS + t + \eta, \quad \eta \sim N(0, \sigma_{2d}^2 I_{2L_0}) \quad (2)$$

where η denotes an isotropic observation noise in the image space; σ_{2d} is the standard deviation, which is dynamically decided according to the variation of the shape in each step; $P = P_{2L_0 \times 3L} = (I_{L_0}, 0)_{L_0 \times L} \otimes P_0$ is the projection matrix with $P_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ and \otimes is the *Kronecker product*; f is the scale parameter; $R = R_{3L \times 3L} = I_L \otimes R_0$ is the rotation matrix and $t = 1_{L_0} \otimes t_0 = 1_{L_0} \otimes (t_x, t_y)'$ is the *translation parameter*. Denote c as the pose parameters $\{\alpha, \beta, \gamma, f, t_x, t_y\}$.

2.4 Parameter Estimation

It is difficult to infer the shape parameter s and pose parameter c from the given 2D shape s_{2d} directly. With the hidden data S , the EM algorithm can be applied

to conduct parameter optimization. Define the Q-function as:

$$\begin{aligned}
 Q(s, c, s^{old}, c^{old}) &= E [\ln P(s, c|s_{2d}, S)|s_{2d}, s^{old}, c^{old}] \\
 &= \int \ln P(s, c|s_{2d}, S) \cdot P(S|s_{2d}, s^{old}, c^{old})dS
 \end{aligned} \tag{3}$$

E-Step: With simple computation from Eqn (1) and (2), we have

$$\begin{aligned}
 -2 \ln P(s, c|s_{2d}, S) &= \frac{1}{\sigma_{3d}^2} \| S - U \cdot s - \bar{S} \|^2 + s' \Lambda^{-1} s \\
 &\quad + \frac{1}{\sigma_{2d}^2} \| s_{2d} - PfRS - t \|^2 + c_1 \\
 -2 \ln P(S|s_{2d}, s^{old}, c^{old}) &= \frac{1}{\sigma_{3d}^2} \| S - U \cdot s^{old} - \bar{S} \|^2 \\
 &\quad + \frac{1}{\sigma_{2d}^2} \| s_{2d} - MS - t \|^2 + c_2
 \end{aligned} \tag{4}$$

where c_1, c_2 are constants and Λ is a diagonal matrix with diagonal elements as leading eigenvalues. The conditional probability $P(S|s_{2d}, s^{old}, c^{old})$ obeys the following Gaussian distribution:

$$P(S|s_{2d}, s^{old}, c^{old}) \sim N(\mu, \Sigma) \tag{5}$$

where ($M = Pf^{old}R^{old}$)

$$\mu = \langle S \rangle = (\sigma_{3d}^{-2}I + \sigma_{2d}^{-2}M'M)^{-1} \cdot [\sigma_{3d}^{-2}(U \cdot s^{old} + \bar{S}) + \sigma_{2d}^{-2}M'(s_{2d} - t^{old})] \tag{6}$$

$$\Sigma = (\sigma_{3d}^{-2}I + \sigma_{2d}^{-2}M'M)^{-1} \tag{7}$$

where $\langle S \rangle$ denotes the conditional expectation $E [S|s_{2d}, s^{old}, c^{old}]$, then we have:

$$\langle SS' \rangle = \Sigma + \langle S \rangle \langle S' \rangle \tag{8}$$

On the other hand, Σ is the inversion of a very large matrix, which is computed expensively. In fact, M has simple form with $M_0 = P_0fR_0$ being a 2×3 matrix.

$$M = (I_{L_0}, 0)_{L_0 \times L} \otimes M_0 \tag{9}$$

Then,

$$\Sigma = \begin{pmatrix} I_{L_0} & 0 \\ 0 & 0 \end{pmatrix}_L \otimes (\sigma_{3d}^{-2}I_3 + \sigma_{2d}^{-2}M'_0M_0)^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & I_{L-L_0} \end{pmatrix}_L \otimes \sigma_{3d}^2I_3 \tag{10}$$

which is much more simple and we only need to compute the inversion of a 3×3 matrix. With the Eqn (5)-(10), the problem is equal to

$$\min_{s,c} \left\langle \frac{1}{\sigma_{3d}^2} \| S - U \cdot s - \bar{S} \|^2 + s' \Lambda^{-1} s + \frac{1}{\sigma_{2d}^2} \| s_{2d} - PfRS - t \|^2 \right\rangle \tag{11}$$

M-Step: Notice that pose parameter c is independent to the shape parameter s . Thus they can be optimized separately.

1) **Optimize shape parameter s :** shape parameter s can be easily derived by setting the derivative of the Q -function to zero:

$$s = \Lambda(\Lambda + \sigma_{3d}^2 I)^{-1} U'(\langle S \rangle - \bar{S}) \tag{12}$$

2) **Semi-closed-form solution for pose parameter c using Quaternion:** From (11),

$$c = \underset{c}{\operatorname{arg\,max}} Q(s, c, s^{old}, c^{old}) = \underset{c}{\operatorname{arg\,min}} \sum_{i=1}^{L_0} \langle \| s_{2d}^i - M_0 S^i \|^2 \rangle \tag{13}$$

where s_{2d}^i denotes the i th point of s_{2d} , and S^i denotes the correspondent point in S . It's a nonlinear optimization problem and can not be optimized directly. Traditionally, unit quaternion [1][2] based pose representation was applied to solve 3D-to-3D pose parameter variation problem. In the following, we will introduce a semi-closed-from algorithm in terms of unit quaternion for pose estimation.

A quaternion is represented as $\overset{\circ}{q} = q_0 + q_x i + q_y j + q_z k$, its complex conjugate is defined as $\overset{\circ}{q}^* = q_0 - q_x i - q_y j - q_z k$ and $S\{\overset{\circ}{q}\} = (q_x, q_y, q_z)'$. A 3D point p is represented by the purely imaginary quaternion $\overset{\circ}{p} = 0 + p_x i + p_y j + p_z k$ and a rotation of p is defined as $\overset{\circ}{q} \overset{\circ}{p} \overset{\circ}{q}^*$, then $f = \overset{\circ}{q} \cdot \overset{\circ}{q}^*$ and $f R p = S\{\overset{\circ}{q} \cdot \overset{\circ}{p} \cdot \overset{\circ}{q}^*\}$. The detailed relation between rotation matrix R_0 , scale parameter f and quaternion $\overset{\circ}{q}$ is referred to [2]. With quaternion representation, the objective function in Eqn (13) can be rewritten as:

$$\min E^2 = \langle \sum_{i=1}^n (\tilde{s}_{2d}^i - S\{\overset{\circ}{q} \overset{\circ}{S}^i \overset{\circ}{q}^*\} - t)' W_i (\tilde{s}_{2d}^i - S\{\overset{\circ}{q} \overset{\circ}{S}^i \overset{\circ}{q}^*\} - t) \rangle \tag{14}$$

where 3D point \tilde{s}_{2d}^i is extended from s_{2d}^i with z -value being zero and W_i represents the directional constraint of the i -th point $W_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ here.

Assume that we have some estimation of $\overset{\circ}{q}$ available at the r -th iteration as $\overset{\circ}{q}_r$, a new estimation $\overset{\circ}{q}_{r+1} = \overset{\circ}{q}_r + \overset{\circ}{\delta}$, then

$$S\{\overset{\circ}{q}_{r+1} \overset{\circ}{S}^i \overset{\circ}{q}_{r+1}^*\} = S\{\overset{\circ}{q}_r \overset{\circ}{S}^i \overset{\circ}{q}_r^* + \overset{\circ}{\delta} \overset{\circ}{S}^i \overset{\circ}{q}_r^* + \overset{\circ}{q}_r \overset{\circ}{S}^i \overset{\circ}{\delta}^* + \overset{\circ}{\delta} \overset{\circ}{S}^i \overset{\circ}{\delta}^*\} \tag{15}$$

Assume $\overset{\circ}{\delta}$ is small with respect to $\overset{\circ}{q}_r$, then Eqn (15) can be approximated as

$$S\{\overset{\circ}{q}_r \overset{\circ}{S}^i \overset{\circ}{q}_r^* + \overset{\circ}{\delta} \overset{\circ}{S}^i \overset{\circ}{q}_r^* + \overset{\circ}{q}_r \overset{\circ}{S}^i \overset{\circ}{\delta}^*\} = f_r R_r \overset{\circ}{S}^i + G_i \overset{\circ}{\delta} \tag{16}$$

where G_i can be derived from the definition.

Denote $v = (q_0, q_x, q_y, q_z, t_x, t_y)'$, $z_i = \tilde{s}_{2d}^i - f_r R_r S^i$ and $G_{vi} = (G_i, (I_{2 \times 2}, 0)')$, we have:

$$\min E^2 = \left\langle \sum_{i=1}^n (z_i - G_{vi}v)' W_i (z_i - G_{vi}v) \right\rangle \quad (17)$$

The optimal solution can be obtained by solving the following traditional function:

$$\sum_{i=1}^n \langle g'_{ij} W_i \sum_{k=1}^6 g_{ik} v_k \rangle = \sum_{i=1}^n \langle g'_{ij} W_i z_i \rangle \quad (1 \leq j \leq 6) \quad (18)$$

where g_{ij} is the j -th column of matrix G_{vi} . G_{vi} is a linear function of S^i , so are g_{ij} and z_i . Therefore both sides of Eqn (18) which are quadratic functions of S^i at most can be directly computed from Eqn (8).

2.5 Dynamic Correspondence Strategy and Weighted Optimization

Hu's work [3] assumed that the correspondences between the contour points and the 3D face model vertexes are known and fixed, which is inappropriate in the case of out-of-plane rotation. Here we assume that the eyes, mouth, and nose points can be matched accurately from 2D to 3D. For the contour points, the absolute value of z coordinate of the normal direction is small. We utilize the information for the contour points and search for more "proper" points to replace the original contour points after iteration, which results in a more precise correspondence between the contour points of 2D image and 3D face vertexes. The comparison between dynamic correspondence and static correspondence is shown in Fig. 1.

Moreover, there will be part of face occluded in a side-view face image. Thus for the occluded points, the location precision will be degraded. We set the

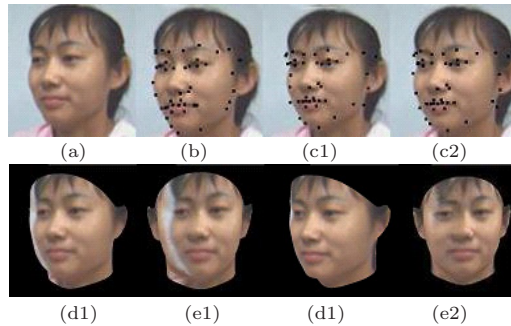


Fig. 1. Comparison between dynamic correspondence and static correspondence. (a) input image; (b) 2D alignment; (c1) 3D geometry reconstruction with static correspondence (black points are the corresponding feature points in 3D model matching the feature points in 2D image in (b).); (c2) 3D geometry reconstruction with dynamic correspondence; (d1)(e1) two views of 3D model with static correspondence; (d2)(e2) two views of 3D model with dynamic correspondence.



Fig. 2. Comparison of the original images with three different views of reconstructed models of various people.

direction constraint W_i of the contour points dynamically, which improves the final result. After the 3D face geometry is reconstructed, the 2D image is mapped to the 3D geometry to generate the texture. Mostly, there are some vertices occluded in the 3D surface; the “mirror” and “interpolate” strategies are applied to improve the reality.

3 Experiments

We constructed a fully automatic 3D face synthesis system based on the proposed algorithm. Our system is fully automatic. The only input is one face image in arbitrary view and there is no user interaction in the whole process.

In our experiments, we used face images with various poses to automatically construct the personalized 3D faces. Fig. 2 shows some experimental results. It shows that our algorithm can reconstruct the 3D face models for different persons in different views; and the generated virtual faces in different views indicate the realistic of the reconstructed 3D model. The faces in the original images in Fig. 2 are in different illumination conditions and there are different skin colors too. One can see the effective 3D reconstruction results.

The whole process to construct a head model from a face image costs less than 1.6 seconds on a PC with Pentium(R) IV 2.8 GHz processor, which is about eighty times faster than the 3D face reconstruction processing [10], ten times faster than Zhang [7], and 1.25 times faster than Hu [3]. The time cost in 3D face geometry reconstruction process is about 0.6 second and it is much faster than Vetter’s [10] method.

4 Conclusions and Future Work

We have proposed a novel framework to construct 3D face model from a single face image in arbitrary view. The experiments show the efficiency and effectiveness of our proposed algorithm. Compared with other related works, its highlights are two-folds: 1) it is fully automatic and handles face images in arbitrary view; and 2) the efficiency and robustness are guaranteed via the EM algorithm integrated with the unit quaternion based pose representation, dynamic correspondence strategy and weighed optimization method.

The efficient 3D face reconstruction with an arbitrary view face image has many applications including 3D model based multi-view face recognition, face pose estimation and virtual reality in 3D game. Currently, we are exploring to efficiently reconstruct the personalized 3D face model based on multiple face images in different views and conduct the face recognition in variant poses; moreover, we are also applying pose estimation results to detect and locate the attention area.

References

1. A.Hill, T.F.Cootes, C.J.Taylor. "Active Shape Models and the shape approximation problem." *Image and Vision Computing*. 14 (8) Aug. 1996 pp 601-608.
2. B.K.P.Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*,4(4):629-642, Apr.1987.
3. Y. X. Hu, D. L. Jiang, S. C. Yan, Lei Zhang, H.J. Zhang. "Automatic 3D Reconstruction for Face Recognition", In FG2004 Proceedings, pages 843-848, 2004.
4. Kin-Man Lam and Hong Yan, An Analytic-to-Holistic Approach for Face Recognition Based on a Single Frontal View, PAMI98, Vol2, No7, page 673-686.
5. J. P. Lewis. Algorithms for solid noise synthesis. In SIGGRAPH '89 Conference proceedings, pages 263-270. ACM, 1989.
6. H. Li, S.C. Yan, L.Z. Peng. "Robust Multi-view Face Alignment with Edge Based Texture", submitted to *Journal of Computer Science and Technology*, 2004.
7. Liu, Z., Zhang, Z., Jacobs, C. and Cohen, M. (2000). Rapid modeling of animated faces from video, Proc. 3rd International Conference on Visual Computing, Mexico City, pp. 58-67. Also in the special issue of *The Journal of Visualization and Computer Animation*, Vol.12, 2001.
8. F.I. Parke. Computer generated animation of faces. In ACM National Conference. ACM, November 1972.
9. F.I. Parke. A Parametric Model of Human Faces. PhD thesis, University of Utah, Salt Lake City, 1974.
10. S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *Computer Vision - ECCV'02*, volume 4, pages 3-19, 2002.
11. N.Magneneat-Thalman, H.Minh,M. Angelis, and D. Thalmann. Design, transformation and animation of human faces. *Visual Computer*, 5:32-39, 1989.
12. M. Tipping and C. Bishop. "Probabilistic principal component analysis" Technical Report Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, Birmingham, UK, September 1997.
13. Ruo Zhang, Ping-Sing Tai, James Edwin Cryer, Mubarak Sha, Shape From Shading: A Survey, *IEEE Trans. On PAMI*, 21(8). pp690-706. 1999

Recognition and Retrieval of Face Images by Semi-supervised Learning

Kohei Inoue and Kiichi Urahama

Kyushu University, Fukuoka-shi, 815-8540 Japan,
{k-inoue,urahama}@design.kyushu-u.ac.jp

Abstract. A semi-supervised learning algorithm based on regularization on graphs is presented and is applied to recognition and retrieval of face images. In a learning phase, the value of classification function is fixed at labeled data and that of unlabeled data is estimated by a regularization scheme whose solution is computed with iteration methods. In a classification phase, the value of classification function of a new datum is computed directly from those of learning data without iterations. The classification rate of the present method is higher than that of the conventional methods such as the basic nearest neighbor rule and the eigenface method. Similarity search of data is also a particular case of the semi-supervised learning where a query is labeled and all data in a database are unlabeled. The relevance degree of data in the database is calculated with regularization and some data with high relevance degree are outputted. The precision of this retrieval scheme is higher than that of the basic similarity search methods.

1 Introduction

In pattern recognition, manual labeling of all learning data is laborious. Hence semi-supervised learning is practically useful, where a new datum is classified on the basis of learning data few of which are labeled and remaining many data are unlabeled[1]. In this paper, we present a semi-supervised learning scheme based on regularization on undirected graphs. Zhu et al.[2] presented a regularization method using an unnormalized graph Laplacian, however, if the structure of clusters in data are skewed then the Laplacian is preferred to be normalized. Zhou et al.[3] presented a regularization method using a normalized Laplacian, however, they formulated the regularization by a function approximation weakly constrained by supervised data, hence the formulation cannot be generalized to a new datum. We formulate in this paper the regularization by function approximation strictly constrained by supervised data, which can be generalized to a new datum. In a learning phase, the value of classification function is fixed at labeled data and that of unlabeled data is estimated by a regularization scheme whose solution is computed with iteration methods. In a classification phase, the value of classification function of a new datum is computed directly from those of learning data without iterations. We verify the higher classification rate of

the present method than the conventional methods such as the nearest neighbor rule and the eigenface method.

We then apply the present semi-supervised learning method to similarity retrieval of data. The semantic gap is a difficult problem in similarity search of multimedia data[4]. It is mainly attributed to a complex structure of clusters in data. Therefore some semantic retrieval methods utilizing clustering of data have been proposed[5]. The clustering of complexly structured data is however difficult in general. We present a similarity retrieval method by employing semi-supervised learning which needs no explicit clustering of data. We verify its higher precision than the basic nearest neighbor search with experiments of retrieval of face images.

2 Function Approximation in Euclidean Space

We start by reviewing a regularization method[6] on Euclidean spaces which is the base of the regularization on graphs. Let the value of a function f is given at a boundary $T \subset \Omega$ in a subspace $\Omega \subset R^m$. T is possible to be only one point. Let the value of the function at a boundary point $x \in T$ be $t(x) \in R$, and an approximated value $a(x)$ is given at every interior point $x \notin T$. We estimate the value of the function $f(x)$ at every interior point $x \notin T$ under this boundary condition, that is, we construct a smooth function f which satisfies $f(x) = t(x)$ at $x \in T$ and minimizes $|f(x) - a(x)|$ over $x \notin T$.

In the standard regularization method[6], f at $x \notin T$ is given by

$$\min_f \int_{x \notin T} [(f - a)^2 + \lambda \|\nabla f\|^2] dx \tag{1}$$

where $\nabla f = [\partial f / \partial x_1, \dots, \partial f / \partial x_m]^T$ is a gradient vector of f . The Euler-Lagrange equation of (1) is

$$f - a - \lambda \Delta f = 0 \tag{2}$$

where $\Delta f = \partial^2 f / \partial x_1^2 + \dots + \partial^2 f / \partial x_m^2$ is the Laplacian of f .

In numerical computation, these equations are discretized, for instance, when $m = 2$ as in image processing, 2-dimensional image plane x is discretized into square grids. If we denote the value of f at a grid point (i, j) by f_{ij} , then (1) is discretized as

$$\min_f \sum_{i,j \notin T} [(f_{ij} - a_{ij})^2 + \lambda \|\nabla f_{ij}\|^2] \tag{3}$$

where $f = \{f_{ij}\}; i, j \notin T$, and the gradient vector is

$$\nabla f_{ij} = \left[\frac{f_{ij} - f_{i-1,j}}{2}, \frac{f_{ij} - f_{i,j-1}}{2} \right]^T \tag{4}$$

The Laplacian is also discretized as

$$\Delta f_{ij} = \sum_{k,l} w_{ij,kl} (f_{kl} - f_{ij}) \tag{5}$$

where $w_{ij,kl} = 1/4$ at the four nearest neighbor points $(k, l) = \{(i - 1, j), (i + 1, j), (i, j - 1), (i, j + 1)\}$ and $w_{ij,kl} = 0$ at other grid points. It holds that $\sum_{k,l} w_{ij,kl} = 1$. Differentiating (3) by f_{ij} and setting it zero, then we get

$$f_{ij} - a_{ij} + \lambda(f_{ij} - \sum_{k,l} w_{ij,kl} f_{kl}) = 0 \tag{6}$$

which coincides with (2) discretized and substituted with (5). In (6), $\sum_{k,l} w_{ij,kl} f_{kl} = \sum_{k,l \notin T} w_{ij,kl} f_{kl} + \sum_{k,l \in T} w_{ij,kl} t_{kl}$.

Equation (6) is rewritten as

$$f_{ij} = \mu \sum_{k,l} w_{ij,kl} f_{kl} + (1 - \mu) a_{ij} \tag{7}$$

where $\mu = \lambda/(1 + \lambda)$. This equation reveals that f_{ij} is a harmonic function with boundary values a_{ij} and t_{ij} , because $\mu \sum_{k,l} w_{ij,kl} + 1 - \mu = 1$. Harmonic functions take their minimum or maximum only at boundaries, hence the following inequalities hold:

$$l \leq f_{ij} \leq u \tag{8}$$

where $l = \min\{\min\{a_{ij}\}, \min\{t_{ij}\}\}$ and $u = \max\{\max\{a_{ij}\}, \max\{t_{ij}\}\}$.

Iterative methods are popularly used for solving (6). The simplest Gauss-Jacobi scheme is written as

$$f_{ij}^{(\xi+1)} = \mu \sum_{k,l \notin T} w_{ij,kl} f_{kl}^{(\xi)} + g_{ij} \tag{9}$$

where $f_{ij}^{(\xi)}$ is the ξ -th iterant and $g_{ij} = \mu \sum_{k,l \in T} w_{ij,kl} t_{kl} + (1 - \mu) a_{ij}$. The coefficient matrix $I - \mu W$ of the linear equation system (7) is diagonally dominant M-matrix[7], hence $f^{(\xi)}$ converges monotonically to the solution of (7) starting from an arbitrary initial value of $f^{(0)}$.

3 Function Approximation on Graphs

Let $i = 1, \dots, n$ be nodes in an undirected graph and $s_{ij} (= s_{ji}), s_{ii} = 0$ be a weight of edge between nodes i and j . The weight s_{ij} is proximity between nodes i and j , for instance, $s_{ij}=1$ if nodes i and j are connected with an edge, otherwise $s_{ij} = 0$.

If we denote a directed edge from node i to j by e_{ij} , a gradient vector of a function f on the graph at the node i is defined by[8]

$$\nabla f_i = [\frac{\partial f_i}{\partial e_{i1}}, \dots, \frac{\partial f_i}{\partial e_{in}}]^T \tag{10}$$

where $\partial f_i / \partial e_{ij}$ is the gradient of f along the edge e_{ij} :

$$\frac{\partial f_i}{\partial e_{ij}} = \sqrt{s_{ij}} (\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}) \tag{11}$$

where $d_i = \sum_{j=1}^n s_{ij}$. For a graph of square grids in image processing, $s_{ij,kl} = 1$ for only the four neighbors, hence $d_i = 4$ and (10) with (11) reduces to (4).

The Laplacian of the graph is defined by[8]

$$\Delta f_i = \frac{1}{\sqrt{d_i}} \sum_{j=1}^n s_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right) = (Lf)_i \tag{12}$$

where $f = [f_1, \dots, f_n]^T$, $L = [l_{ij}]$; $l_{ii} = 1$, $l_{ij} = -s_{ij}/\sqrt{d_i d_j}$ ($i \neq j$), and $(Lf)_i$ is the i -th element of Lf . L is the normalized Laplacian matrix of the graph:

$$L = I - D^{-1/2} S D^{-1/2} = D^{-1/2} (D - S) D^{-1/2} \tag{13}$$

where I is the unit matrix and $S = [s_{ij}]$, $D = \text{diag}(d_1, \dots, d_n)$. Equation (12) reduces to (5) for a graph of square grids with sign inversion which comes from that the sign of the graph Laplacian in the definition (12) is reverse to the Euclidean Laplacian.

3.1 Standard Regularization

Let us be given a function value t_i at a boundary point $i \in T$ and an approximate value a_i at all the remaining interior points $i \notin T$. Equation (3) is generalized to graphs as

$$\min_f \sum_{i \notin T} [(f_i - a_i)^2 + \lambda \|\nabla f_i\|^2] \tag{14}$$

Substituting (10) with (11) into the second term, we get

$$\min_f \sum_{i \notin T} [(f_i - a_i)^2 + \lambda \sum_{j=1}^n s_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2] \tag{15}$$

Differentiating this objective function with f_i and equating it to zero, we obtain

$$f_i - a_i + \lambda \frac{1}{\sqrt{d_i}} \sum_{j=1}^n s_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right) = 0 \tag{16}$$

With reference to (12), (16) is rewritten as

$$f_i - a_i + \lambda \Delta f_i = 0 \tag{17}$$

which corresponds to (2) with the sign inversion as the same as in the Laplacian.

If we transform the variables as $\tilde{f}_i = f_i/\sqrt{d_i}$, $\tilde{a}_i = a_i/\sqrt{d_i}$, $\tilde{t}_i = t_i/\sqrt{d_i}$, (16) is written as

$$d_i(\tilde{f}_i - \tilde{a}_i) + \lambda \sum_{j=1}^n s_{ij}(\tilde{f}_i - \tilde{f}_j) = 0 \tag{18}$$

Hence \tilde{f} is a harmonic function on the graph, and satisfies inequalities similar to (8), which is written with the original variables as

$$l \leq f_i \leq u \tag{19}$$

where $l = \sqrt{d_i} \min\{\min\{a_i/\sqrt{d_i}\}, \min\{t_i/\sqrt{d_i}\}\}$, $u = \sqrt{d_i} \max\{\max\{a_i/\sqrt{d_i}\}, \max\{t_i/\sqrt{d_i}\}\}$.

3.2 Numerical Solution

If we divide the set of j in (16) into $j \notin T$ and $j \in T$, then (16) is written as

$$f_i - a_i + \lambda(f_i - \sum_{j \notin T} w_{ij} f_j - \sum_{j \in T} w_{ij} t_j) = 0 \tag{20}$$

where $w_{ij} = s_{ij} / \sqrt{d_i d_j}$. Equation (20) corresponds to (6). From (20), we get

$$f_i = \mu(\sum_{j \notin T} w_{ij} f_j + \sum_{j \in T} w_{ij} t_j) + (1 - \mu)a_i \tag{21}$$

where $\mu = \lambda / (1 + \lambda)$. The Gauss-Jacobi iteration for (21) is written as

$$f_i^{(\xi+1)} = \mu(\sum_{j \notin T} w_{ij} f_j^{(\xi)} + \sum_{j \in T} w_{ij} t_j) + (1 - \mu)a_i \tag{22}$$

Since the coefficient matrix $I + \lambda L$ in (20) is an M-matrix[7], the iterant $f^{(\xi)}$ converges monotonically to the solution of (20) for an arbitrary initial value $f^{(0)}$.

4 Application to Semi-supervised Pattern Classification

Let us consider classification of test data based on learning data where the value of classification function is known for only few labeled data while is unknown for remaining unlabeled data. We estimate the function value of unlabeled learning data by the above function approximation method and use the whole learning data for classifying a new datum.

4.1 Function Approximation for Learning Data

Let a set of patterns be composed of m classes and let us be given n learning data in which the value $t_{ci}(c = 1, \dots, m)$ of classification function is given only for labeled data $i \in T \subset \{1, \dots, n\}$. If a datum $i \in T$ is a member in a class c_* , then $t_{c_*i} = 1$ and $t_{ci} = 0(c \neq c_*)$. On the basis of these supervised function values, we estimate function value f_{ci} of unlabeled data $i \notin T$. In this case, the approximation value of function is set $a_{ci} = 0$ for all $i \notin T$. Hence (19) becomes in this case as

$$0 \leq f_{ci} \leq \sqrt{d_i} \max\{\frac{1}{\sqrt{d_i}}\} \tag{23}$$

Let us represent the learning data with a complete undirected graph with n nodes where the weight of edges is the proximity between data i and j as expressed by $s_{ij} = e^{-\alpha \|p_i - p_j\|^2}$ where p_i is a feature vector of datum i .

We estimate the function value f_{ci} at unlabeled nodes $i \notin T$ by the method described in section 3. Their computation is executed independently for each

class c , that is, we compute f_{ci} on the basis of $t_{ci}(i \in T), a_{ci}(i \notin T)$ for each c separately. Equation (21) reads in this case

$$f_{ci} = \mu \left(\sum_{j \notin T} w_{ij} f_j + \sum_{j \in T_c} w_{ij} \right) \tag{24}$$

where T_c is a set of learning data included in class c . The initial value of iterative solution scheme is set as $f_{ci}^{(0)} = 0$.

4.2 Classification of Test Data

Test data are classified on the basis of these learning data. When we classify test data, the value of the classification function is already computed for all of the learning data. Hence the classification of a test datum is a particular case of semi-supervised learning where all the learning data are labeled and only one test datum is unlabeled. We estimate the function value of the test datum by above described function approximation method for semi-supervised learning.

Let the feature vector of a test datum be p . With reference to (24), the function value of the test datum p in a class c is given by

$$f_c(p) = \mu \left(\sum_{i \notin T} \frac{s_i(p) f_i}{\sqrt{d(p) d_i}} + \sum_{i \in T_c} \frac{s_i(p)}{\sqrt{d(p) d_i}} \right) \tag{25}$$

where all of i are learning data and the sets T and T_c are the same as those in (24), and $s_i(p) = e^{-\alpha \|p - p_i\|^2}, d(p) = \sum_{i=1}^n s_i(p)$. Since d_i and f_i are the values of learning data which are already computed, we can compute $f_c(p)$ directly from (25) without iterations. We compute (25) for each class c and classify the test datum to the class $c_* = \arg \max_c \{f_c(p)\}$.

4.3 Experiments of Face Recognition

The above method is applied to the identification of persons with their face images. The size of images is 56×46 . The feature vector p_i is an array of luminance of every pixel in an image and $s_{ij} = e^{-\alpha \|p_i - p_j\|^2}$ with $\alpha = 5 \times 10^{-6}$ and $\mu = 0.99$ in (21). The number of 575 face images of 20 persons are used as learning data and other 100 images are used for test. The classification rates are plotted in Fig.1 where the abscissa is the number of labeled images per one person and the ordinate is classification rates. The solid line denotes the present method, the broken line is the rate of the eigenface method and the dotted line is that of the basic nearest neighbor rule. In the eigenface method, the feature vector is projected into 100 dimensional space via the principal component analysis and the data is classified by the nearest neighbor rule in this low dimensional space. The present method always outperforms the conventional methods particularly when the number of labeled data is small owing to label propagation effect of the regularization method.

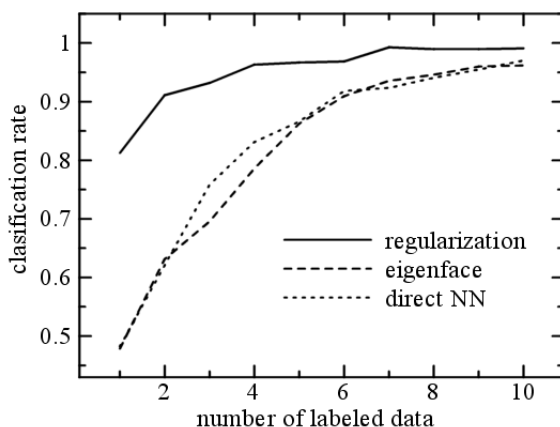


Fig. 1. Classification rates

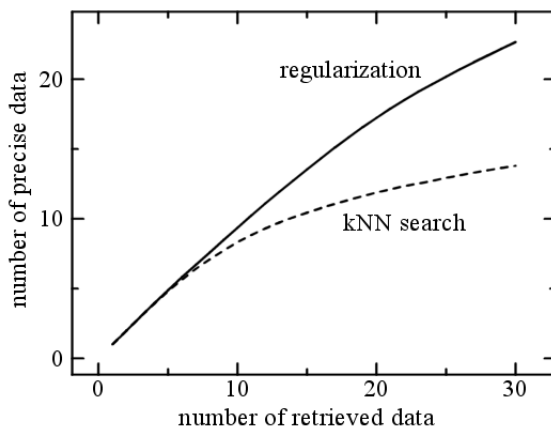


Fig. 2. Precision of similarity search

5 Application to Similarity Search

Similarity search of data is also a particular case of semi-supervised learning for pattern classification where the number of classes is two, i.e., data relevant to a query and other irrelevant data. The function is the degree of relevance of each datum to a query whose function value is 1 while function values are unknown for all data in a database. The conventional k NN search using the distance directly is a particular case where every similarity between database data is fixed as $s_{ij} = 0$.

5.1 Function Approximation

We set the whole data including database data and a query as a learning dataset and compute function values of database data by the above regularization method. We then output some database data in the order of the computed function values.

5.2 Experiments of Face Image Retrieval

We have experimented similarity retrieval of face images from the dataset used in the above experiments of face recognition. Similarity between face images are the same as s_{ij} in the above classification experiments. The number of images in the retrieved ones in the same class as a query image is plotted in Fig.2 where the solid line denotes the present method and the broken line is the result of direct k NN search based on the Euclidean distance between a query and database data. The retrieval precision of the present method is higher than the conventional similarity search method. Hence the present retrieval method is effective for narrowing the semantic gap in the retrieval of multimedia data.

6 Conclusion

We have presented a semi-supervised learning method based on the regularization on graphs and have applied it to classification and retrieval of face images. The present method outperforms the conventional methods, however, demands longer computational time than the conventional methods because the present method employs iterative solution methods. Acceleration by using such as the Gauss-Seidel or SOR iteration schemes[7] is under study.

References

1. Seeger, M.: Learning with labeled and unlabeled data. Tech. Report, Univ. Edinburgh (2002)
2. Zhu, X., Ghahramani, Z. and J. Lafferty: Semi-supervised learning using gaussian fields and harmonic functions. 20th Int. Conf. Machine Learning (2003)
3. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. and Scholkopf, B.: Learning with local and global consistency. Adv. Neural Inf. Process. Syst. **16** (2004)
4. Zhao, R. and Grosky, W. I.: Negotiating the semantic gap: from feature maps to semantic landscapes. Patt. Recogn. **35**, 3 (2002) 593-600
5. Chen, Y., Wang, J. Z. and Krovetz, R.: Content-based image retrieval by clustering: 5th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval. (2003) 193-200
6. Tikhonov, A. N. and Arsenin, V. Y.: Solution of ill-posed problems. Wiley, New York (1977)
7. Ortega, J. M.: Numerical analysis: a second course. Society Indust. & Applied Math. (1990)
8. Chung, F. R. K.: Spectral graph theory. Amer. Math. Society (1997)

3-D Facial Expression Recognition-Synthesis on PDA Incorporating Emotional Timing

Doo-Soo Lee, Yang-Bok Lee, Soo-Mi Choi, Yong-Guk Kim, and
Moon-Hyun Kim

School of Computer Engineering, Sejong University, Seoul, Korea
{smchoi, ykim, mhkim}@sejong.ac.kr

Abstract. This paper describes a pipeline by which facial expression of the face is recognized and then 3-D facial animation is synthesized in the remote place based upon timing information of the facial. The system first detects a facial area within the given image and then classifies its facial expression into 7 emotional weightings. Such weighting information, transmitted to the PDA via a mobile network, is used for non-photorealistic facial expression animation. A cartoon-like shading is used to render a 3-D avatar that conveys a familiar and yet unique facial character, even without employing extensive polygons. It turns out that facial expression animation using emotional curves is more effective in expressing the timing of an expression comparing to the linear interpolation method. The emotional avatar embedded on a mobile platform has some potential in conveying emotion between peoples in Internet.

1 Introduction

We identify someone by looking at his/her face, since each person typically has unique and distinctive features in the face. Moreover, human face is a great communication device, because the face can evoke diverse facial expressions according to the internal emotions. So one can read someone's emotional state from his/her facial expression and respond to it appropriately.

Although the study on facial expression has a rather long history since Charles Darwin, automatic analysis of human facial expressions using the computer is a recent trend [5]. It is known that there are six basic (or prototypical) facial expressions for humans across the diverse ethnicities and cultures: happiness, surprise, sadness, fear, anger and disgust [3]. Our present study is also based upon this assumption and deals with such cases.

Humans are able to recognize facial expressions of the rendered face on the screen (or paper) as well as of the real face. In fact, researchers in the computer graphics and multimedia areas have been developed a series of face models and their implementations for using in diverse human-computer interaction applications or for animating an avatar in the cyberspace. As an example, the early attempt of unifying the facial expression recognition and facial animation envisioned such virtual face avatar as a future communication media [5]. In fact, such

avatar has some advantage compared to the real face since it can be anonymous, friendly, funny and animated in real-time base [8].

The present study will present a multi-step pipeline: in the first step the facial expressions of the face within the video images are recognized as a time sequence; in the second step the extracted emotional information is transmitted to a remote client via the mobile network; in the final step the cartoon-like 3-D facial expressions are rendered on a PDA using that information.

Despite the fact that the PDA (and smart-phone) become widespread because of its mobility and convenience as the post-PC systems, its processing power for rendering the 3-D animation is yet limited [9]. To circumvent such shortcoming, we have designed a non-realistic 3-D model that does not need many polygons in drawing a face.

The structure of the paper is as follow. In section 2, we describe the automatic recognition of facial expressions. The face-muscle model and non-photorealistic rendering will be described in section 3. In section 4, we show how to incorporate timing aspect of emotion into animation. Experimental results and conclusions will be presented in section 5.

2 Automatic Facial Expression Recognition

The pipeline proposed in this paper consists of two main parts. The first part classifies its facial expression of the face contained in the video from the camera and the system is operated in the server. The second part is within the client such as a PDA where a 3D face is animated according to the emotion graph transmitted from the server.

In general, the facial expression recognition system consists of three stages: in the first stage, it detects a face area within the given image, called the face detection; in the second stage, the positions of three facial features (i.e. two eyes and mouth) will be located within the detected face for normalizing the face area; the final stage classifies a facial expression of the given face using a classifier.

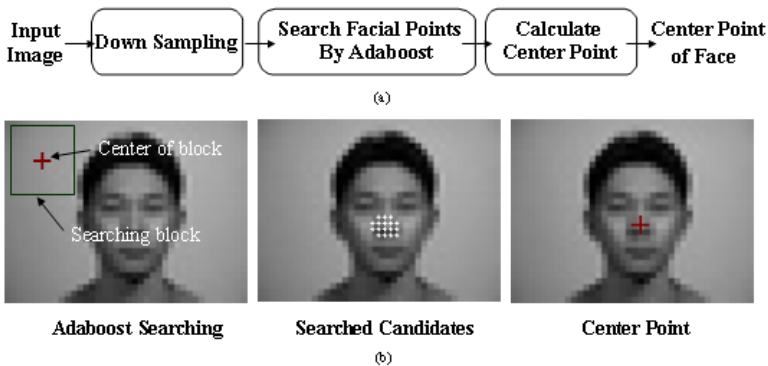


Fig. 1. Illustration of finding a face within an image using Adaboost algorithm.

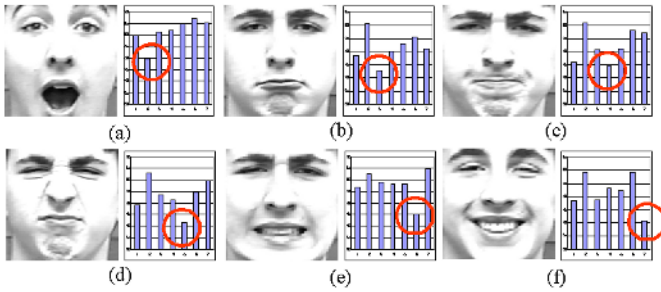


Fig. 2. The results of automatic facial expression recognition.

In our study, Adaboost algorithm is used for the face detection task as shown in Fig. 1(a). In addition, we can train the algorithm to deal with size variation and rotation of the face. In any case, once a face is detected using this machine, the face image needs to be normalized by locating the positions of two eyes and the mouth. The same algorithm is used for this task as well. Note that the location of the mouth is included for the present case, which contrasts to the face perception case where only the locations of two eyes are typically used for normalization.

After the normalization stage, Gabor wavelets are applied to the 20X20 grid drawn over the image, and the convolution output from this operation is forwarded to the next stage to classify the facial expression. For this purpose, we have used the EFM (Enhanced Fisher Discriminant Model), which in fact combines PCA (Principal Component Analysis) with Fisher Discriminant, and it was initially developed for the face perception. Comparative study shows that performance of the system has increased by adding Gabor wavelets at least 10% and the EFM outperforms the PCA. When the Gabor wavelets and the EFM are combined, the recognition rate of the system reaches to 92%.

Fig. 2 shows six basic facial expressions of a subject from the standard Cohn-Kanade database [4], and the corresponding emotional similarity measurements performed by EFM. Each facial image are shown as a bar graph beside each image. In the bar graph, each vertical bar represents the similarity measure to one of the seven emotions from neutral (bar1), surprise (bar2), fear (bar3), sadness (bar4), anger (bar5), disgust (bar6) and happiness (bar7) from the left to the right, respectively. For instance, as Fig. 2(a) is the image for ‘surprise’, the height of the bar2 is the shortest among seven vertical bars, indicating that the measurement by the system agrees with our visual perception.

3 A Muscle-Based Face Model and Non-photorealistic Rendering

In order to increase the speed of animation on the PDA, we have developed a simple muscle-based 3-D face model. The muscle movements for animating a 3-D

face are mainly based on Waters' linear muscle model [6,7]. And the muscle control parameters in the model are based on the FACS (Facial Action Coding System)[3], which can describe all possible basic 'Action Units' within a human face.

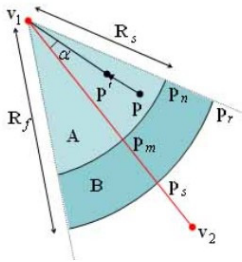
A muscle is modeled as the vector that has a direction from the point of bone \mathbf{v}_1 to the point of skin \mathbf{v}_2 as illustrated in Fig. 3(a). It is assumed that there is no displacement at the bony attachment and maximum deflection occurs at the point of the skin. The extension of vector field is described by cosine functions and fall off factors that has a cone shape. A dissipation of the force is passed to the adjoining tissue from the sector \mathbf{A} to \mathbf{B} . \mathbf{R}_s and \mathbf{R}_f represent the fall-off radius start and its finish, respectively. The new displacement \mathbf{p}' of an arbitrary vertex \mathbf{p} within the zone $(\mathbf{v}_1, \mathbf{p}_r, \mathbf{p}_s)$ is computed as follows:

$$\mathbf{p}' = \mathbf{p} + \cos(\alpha)kr(\mathbf{p}\mathbf{v}_1 / \|\mathbf{p}\mathbf{v}_1\|) \tag{1}$$

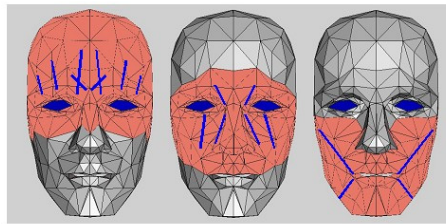
where α is the angle between the vectors $(\mathbf{v}_1, \mathbf{v}_2)$ and $(\mathbf{v}_1, \mathbf{p})$, \mathbf{D} is $\|\mathbf{v}_1 - \mathbf{p}\|$, k is a fixed constant representing the elasticity of skin, and r is the radial displacement parameter:

$$r = \begin{cases} \cos(1 - \mathbf{D}/\mathbf{R}_s) & \text{for } \mathbf{p} \text{ inside zone } (\mathbf{v}_1, \mathbf{P}_n, \mathbf{P}_m) \\ \cos((\mathbf{D} - \mathbf{R}_s)/(\mathbf{R}_f - \mathbf{R}_s)) & \text{for } \mathbf{p} \text{ inside zone } (\mathbf{P}_n, \mathbf{P}_r, \mathbf{P}_s, \mathbf{P}_m) \end{cases} \tag{2}$$

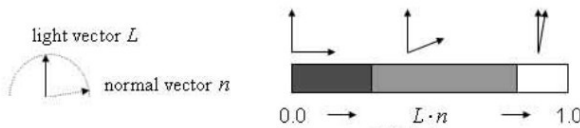
The complexity of the algorithm depends on the number of vertices it has to check to see whether they are inside the zone of influence. To increase the speed, the number of vertices and polygons is reduced except for the expressive regions such as two eyes, the mouth and the nose. The present base key-model has 296 polygons for the whole face and 180 polygons for the eyeball region, respectively. The face model is divided into three regions, i.e. the upper, middle and lower parts of the face as shown in Fig. 3(b).



(a) The muscle model



(b) Sub-regions of the face model with embedded muscles



(c) 1-D texture map

Fig. 3. The muscle-based model (a, b) and the texture map (c) for cartoon-like shading.

Since it is known that a muscle within a sub region of the face will have influence on the others, it is possible to eliminate all the vertices that are outside the region by checking the flags. Opening of the mouth is animated by rotating the vertices of the lower part of the face about a jaw pivot axis. To create a natural oval-shape mouth, the vertices on the lower lip are rotated by different amounts. The upper lip is also affected by the jaw rotation.

In order to animate humors and subtle emotions, our system employs cartoon-like shading [1] as a non-photorealistic rendering technique (See Fig. 7), rather than smoothly interpolating shading across a model as in *Gouraud* shading. In our system, the diffuse lighting at the vertices is defined by the following lighting equation:

$$C_i = a_g a_m + a_l a_m + (\max\{L \cdot n, 0\}) d_l d_m \quad (3)$$

C_i is the vertex color, a_g is the coefficient of global ambient light, a_l and d_l are the ambient and diffuse coefficients of the light source, and a_m and d_m are the ambient and diffuse coefficients of the object's material. L is the unit vector from the light source to the vertex, and n is the unit normal to the surface at the vertex. Instead of calculating the colors per vertex, a 1-D texture map of a minimal number of colors is computed and stored ahead of time (i.e. illuminated main color, shadow color, and highlight color).

4 Timing Aspect of Facial Expressions and Their Representation

The widely used technique in animating faces is a key-frame interpolation. However, recent study in facial behavior suggests that our facial expressions cannot model properly with such simple assumption [2]. We know that one of 12 basic principles for animators published from Disney TM emphasizes the timing aspect of a character animation. For instance, the true intention of a character can be conveyed genuinely, when the timing of the action is properly executed. In fact, it is known that animators adopt a discrete interpolation gimmick by inserting an anticipation frame between two key-frames. Here, Fig. 4 shows such

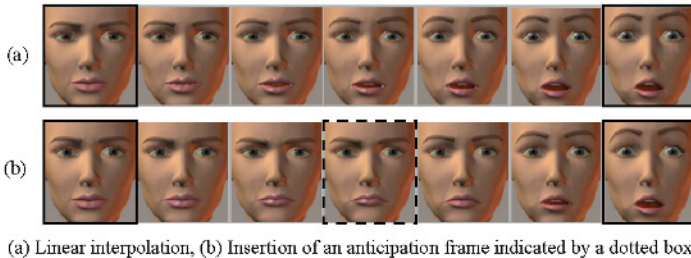


Fig. 4. Two different ways of animating the surprise facial expression.

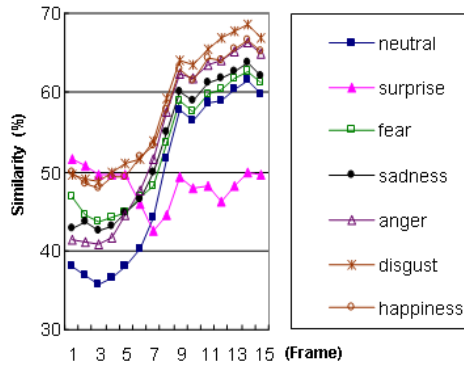


Fig. 5. Emotional curves obtained by the facial expression recognition system.

example, which is generated using Poser 5 TM . In Fig. 4(b), timing sequences of the surprise facial expression contain an abrupt change of facial action.

In this paper, we describe a facial expression recognition-synthesis pipeline in which the emotional curves obtained from the front recognition stage provide the timing information for the facial animation. Fig. 5 shows the seven emotional curves (i.e. neutral, surprise, fear, sadness, anger, disgust, happiness) obtained by automatic facial expression recognition. The emotional curves give weighting information for facial synthesis. The horizontal axis shows the change of time and the vertical one represents the similarity measure. The separation between surprise curve and the other ones indicate that facial expression is changed neutral into surprise at time 7.

As the basis for generating non-linear expression synthesis, we use the concepts of weighting information based on emotional curves and blending functions. Each emotion has an emotional curve obtained by automatic facial expression recognition. Each emotional expression also has an associated target set of facial control parameter values. The actual parameter value used at a given animation frame time $F_p(t)$ is determined by blending emotional curves using a weighted average:

$$F_p(t) = \frac{\sum_{e=1}^n (W_{ep}(t)T_{ep})}{\sum_{e=1}^n W_{ep}(t)} \quad (4)$$

where n is the number of recognized expressions and T_{ep} are the target facial control parameter values. W_{ep} is the weight of an emotional expression. For stable animation, two dominant expressions are blended in runtime.

5 Implementation and Results

The relationship between the original faces and the corresponding non-photorealistic 3-D faces is illustrated in Fig. 6. The PDA used as a client system is an iPAQ 3950 (400MHz, 64Mbytes) from HP, and we have developed the software

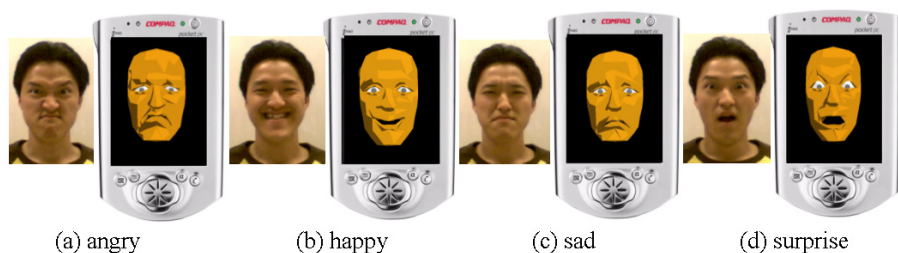


Fig. 6. Four basic facial expressions and their corresponding animations.

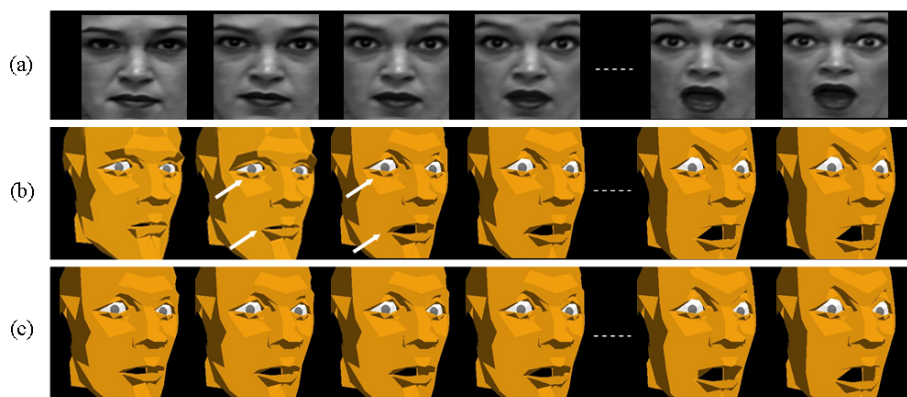


Fig. 7. Comparison between linear interpolation and emotional curve-based animation.

system using Embedded Visual C++ 3.0 and PocketGLTM as a PocketPC 3-D graphics library. Considering the limitation of processing speed of the PDA, we have used minimum number of polygons in rendering the 3-D face, and have adopted an optimization method for the floating-point operations. We were able to animate the 3-D face with a speed of 9~10 fps, by dividing the face into three regions and by applying optimized float-point operation.

Fig. 7 shows the non-photorealistic 3-D faces implemented in the present study. The input face image is in Fig. 7(a). The animation sequence by emotional curves is illustrated in Fig. 7(b), while 'surprise' sequence by linear interpolation method in Fig. 7(c). When you compare two sequences, it is possible to observe that the surprise expression is attained suddenly from frame 2 to frame 3 in Fig. 7(b), whereas the same expression is evolved in a linear fashion in Fig. 7(c).

As the recent research suggests that humans are very sensitive on the timing of facial expression, the facial animation based upon emotional curves reflects more effectively user's emotional state in the time sequence [2]. Moreover, the face contains a small number of polygons, it appears to be smooth and primary facial features such as the mouth and two eyes are distinguished.

6 Conclusions and Discussion

The present study demonstrates a pipeline where the facial expression recognition system is integrated with the 3-D facial animation on the PDA via a mobile network. Both automatic analysis of human facial expression and 3-D facial expression animation are still evolving areas [5,9]. Thus, combining two disciplines has been a non-trivial task [10]. We present a basic framework in which an emotional avatar that attains human emotional states can be generated. In particular, we have found that facial expression synthesis using the emotional curves obtained from the facial expression recognition is more natural in expressing the timing of a facial animation than using the simple linear interpolation method. We have also developed a cartoon shading method as one of the non-photorealistic techniques in rendering an avatar on the PDA without employing extensive polygons.

As many smart phones with a camera are available, and hardware accelerators for 3-D graphics will be included in the near future, it will be possible to exchange messages that contain a 3-D emotional avatar conveying sender's emotional story. This kind of 3-D emotional avatar will obviously replace conventional 2-D emoticons, usually adopted in Internet messenger systems. We expect that this kind of emotional avatar can be an effective communication mean in the ubiquitous computing environment.

References

1. Adam, L., Marshall, C., Marris, M., and Blackstein, M.: Stylized Rendering Techniques for Scalable Real-Time 3D Animation. In Symposium of Non-Photorealistic Animation and Rendering (NPAR) 2000, pp. 13-20, 2000.
2. Cohn, J., et al.: Multimodal Coordination of Facial Action, Head Rotation, and Eye Motion during Spontaneous Smiles. IEEE Conference on Automatic Face and Gesture, Korea, May 2004.
3. Ekman, P., and Friesen, W.: Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues. Palo Alto. Consulting Psychologists Press, 1975.
4. Kanade, T., Cohn, J., and Tian, Y.: Comprehensive Database for Facial Expression Analysis. Proc. Int'l Conf. Face and Gesture Recognition, pp. 46-53, 2000.
5. Thalmann, N., Kalra, P., and Escher, M.: Face to Virtual Face. Proceeding of IEEE, pp.870-883, 1998.
6. Parke, F. I., and Waters, K.: Computer Facial Animation. A K Peters, 1996.
7. Waters, K.: A Muscle Model for Animating Three-Dimensional Facial Expressions. SIGGRAPH'87, Vol. 21, pp.17-24, 1987.
8. Buck, I., Finkelstein, A., Jacobs, C., et al.: Performance-Driven Hand-Drawn Animation. In Symposium of Non-Photorealistic Animation and Rendering (NPAR) 2000, pp. 101-108, 2000.
9. Pandzic, I. S.: Facial Animation Framework for the Web and Mobile Platforms. Web3D, pp. 24-28, 2002.
10. Byun, M., Badler, I.: Qualitative Parametric Modifiers for Facial Animations. SIGGRAPH, pp. 65-71, 2002.

Probabilistic Face Tracking Using Boosted Multi-view Detector

Peihua Li^{1,2} and Haijing Wang³

¹ College of Computer Science and Technology, Heilongjiang University,
Heilongjiang Province, 150001, China

² IRISA/INRIA Rennes, Campus Universitaire de Beaulieu,
35042 Rennes Cedex, France

peihualj@hotmail.com, pli@irisa.fr

³ Dept. of Computer Science and Engineering, Harbin Institute of Technology, China

Abstract. Face tracking in realistic environments is a difficult problem due to pose variations, occlusions of objects, illumination changes and cluttered background, among others. The paper presents a robust and real-time face tracking algorithm. A novel likelihood is developed based on a boosted multi-view face detector to characterize the structure information. The likelihood function is further integrated with particle filter which can maintain multiple hypotheses. The algorithm proposed is able to track faces in different poses, and is robust to temporary occlusions, illumination changes and complex background. In addition, it enjoys a real-time implementation. Experiments with a challenging image sequence shows the effectiveness of the algorithm.

1 Introduction

Face tracking has widespread applications in multimedia analysis, video and communication processing, and human-machine interface. Robust and real-time face tracking is, however, a difficult problem in realistic world, because of pose variations of human faces, partial or complete occlusions of object, illumination changes and complex background, among others. Many researchers have been contriving to develop reliable face tracking algorithms.

Spors and Rabenstein [1] present a real-time face localization and tracking algorithm for color video. The face localization is based on skin color segmentation, and tracking is accomplished through Kalman filtering, which estimates the position and size of face with the help of eye localization based on PCA. Birchfield [2] develops a real-time system which is able to track a person's head to automatically control the camera's pan, tilt and zoom. The head outline is modelled as an ellipse, whose size and position are continually updated by a local search combining the output of a module concentrating on the intensity gradient around the ellipse's perimeter with that of another module focusing on the color histogram of the ellipse's interior. CAMSHIFT algorithm [3] intends to tracking face for use in a perceptual user interface. The algorithm depends on skin color projection onto flesh probability image and on gradient optimization method.

Comaniciu and Ramesh propose an efficient framework for tracking of human faces [4,5]. The tracking is based on mean shift algorithm, which aims at optimization of a metric function between the target and candidate distributions, both represented as multi-channel color histogram.

Most face tracking algorithms above-mentioned depend on color information, which is well known to be sensitive to illumination changes. In addition, complex situations arising from realistic environment, such as pose variations, partial or complete occlusions, illumination changes and cluttered background, often lead visual measurement to be ambiguous and non-linear, making it necessary to turn for robust tracking to advanced algorithms that can maintain multi-hypotheses, instead of Kalman filter-like algorithms which maintain single hypothesis [6,7]. We propose to use structure information of the face, which is captured by a novel likelihood function built upon a boosted multi-view face detector [8]. The structure information achieved in this way is insensitive to illumination changes and visual clutter. The likelihood presented is further combined with particle filter [7], which can well deal with nonlinear, non-Gaussian problems.

The remainder of the paper is structured as follows. Section 2 introduces the generative model for object tracking, after the shape and motion modes being described, the likelihood function being presented based on a boosted multi-view face detector. Section 3 describes the particle filter based face tracking algorithm. Section 4 makes experiments to demonstrate the performance of the algorithm. The concluding remarks are made in section 5.

2 Generative Model for Object Tracking

2.1 Shape Model

The shape of the object is parameterized as a B-spline curve as follows [9]

$$\mathbf{r}(s, t) = \begin{bmatrix} x(s, t) \\ y(s, t) \end{bmatrix} = \begin{bmatrix} \mathbf{B}(s)^T & 0 \\ 0 & \mathbf{B}(s)^T \end{bmatrix} \begin{bmatrix} \mathbf{Q}^x(t) \\ \mathbf{Q}^y(t) \end{bmatrix} \quad (1)$$

where $\mathbf{B}(s) = [b_0(s) \ \cdots \ b_{J-1}(s)]^T$, for $0 \leq s \leq L$, $b_i(s)$ ($0 \leq i \leq J-1$) is the i th B-spline basis function, \mathbf{Q}^x is a column vector consisting of x coordinates of all the control points and so is \mathbf{Q}^y (the time index t is omitted hereafter for simplicity), and L is the number of spans. The configuration of the spline is restricted to a shape-space of vectors \mathbf{X} defined by

$$\begin{bmatrix} \mathbf{Q}^x \\ \mathbf{Q}^y \end{bmatrix} = \mathbf{W}\mathbf{X} + \begin{bmatrix} \bar{\mathbf{Q}}^x \\ \bar{\mathbf{Q}}^y \end{bmatrix} \quad (2)$$

where \mathbf{W} is a shape matrix whose rank is less than $2J$, and $\bar{\mathbf{Q}} = [\bar{\mathbf{Q}}^x \ \bar{\mathbf{Q}}^y]^T$ is a template of the object. Typically the shape-space may allow translation and scale deformation of the template shape and the shape matrix \mathbf{W} has the following form

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & \bar{\mathbf{Q}}^x & 0 \\ 0 & 1 & 0 & \bar{\mathbf{Q}}^y \end{bmatrix} \quad (3)$$

2.2 Dynamical Model

The motion equation of the state in the shape space is modelled as the multi-dimensional second order auto-regression (AR) process [9], which generally can be seen as the discretized form of a continuous stochastic second order dynamic system. This multi-dimensional AR process may be regarded as the direct extension of a 1D AR process, which has the following form

$$x_k = a_1 x_{k-1} + a_0 x_{k-2} + b_0 \nu \quad (4)$$

where $a_1 = -\exp(-2\beta\tau)$, $a_0 = 2\exp(-\beta\tau)\cos(\omega\tau)$, $b_0 = \sqrt{1 - a_1^2 - a_0^2 - 2\frac{a_1 a_0^2}{1 - a_1}}$, ν is one dimensional Gaussian i.i.d. noise, β , ω and τ are the damping coefficient, the oscillation period and the sampling period of the corresponding continuous system. It is desirable, in practice, to model the translation and the shape variations of the contour separately, so the 1D AR process is extended respectively to two complementary subspaces of the shape space: translation subspace and deformation subspace. Then the multi-dimensional motion model can be represented as below

$$\mathbf{X}_k = \mathbf{A}_1 \mathbf{X}_{k-1} + \mathbf{A}_0 \mathbf{X}_{k-2} + \mathbf{B}_0 \mathcal{V} \quad (5)$$

where \mathcal{V} is multi-dimensional Gaussian i.i.d.

2.3 Observation Model

In object tracking, the observation model is responsible for extracting visual information when every new image is available. Our observation model is based on one of novel machine learning algorithm—AdaBoost [10].

Training of a Boosted Multi-view Face Detector. In the face detection area, Viola and Jones [8] first realize the selection of critical visual features from a large set of Harr-like features and the training of Adaboost [10] simultaneously. Thanks to the introduction of a new image representation called “Integral Image”, which allows the features used to be computed efficiently, and combination of weak classifiers in a cascade, which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions, their algorithm is computationally efficient.

One of the main ideas of the AdaBoost algorithm is to maintain a distribution or a set of weights over the training set, by calling a given weak learning algorithm repeatedly in a series of rounds $t = 1, \dots, T$. The weights on training examples on the round k is denoted $D_k(i)$ $i = 1, \dots, N$. Initially, weights are set equally among face and non-face examples. But on each round, the weights of incorrectly classified examples are increased, so that in the later consecutive training stages the weak learner is forced to focus on the hard examples in the training set. More precisely, the weak learning algorithm’s job is, among a set of weak functions, to select one feature $f_{j^*}(\mathbf{x})$ which satisfies Eq. (6). Once the weak learner h_k has

been received, AdaBoost chooses α_k , as described by Eq. (7), which measures the importance that it assigns to h_k . The distribution D_t is then updated using the rule as shown by Eq. (8). The final hypothesis H is an average of the T weak hypotheses. The training algorithm is described briefly as follows.

1. Given a collection of N_1 face and N_2 non-face examples $\{(\mathbf{x}_i, y_i), i = 1, \dots, N\}$, where $N = N_1 + N_2$, $y_i = +1$ denotes face example, $y_i = -1$ denotes non-face example. Set weights $D_1(i) = 0.5N_1, 0.5N_2$ for $y_i = +1, -1$ respectively.
2. For $k = 1, \dots, T$
 - a. For a set of weak functions, $f_j(\mathbf{x}_i), j = 1, \dots, M$, choose the weak function as the k th weak learner $h_k(\bullet) = f_{j^*}(\bullet)$ for which

$$p_{D_k}(f_{j^*}(\mathbf{x}_i) \neq y_i) = \operatorname{argmin}_j p_{D_k}(f_j(\mathbf{x}_i) \neq y_i) \tag{6}$$

- b. Update the weight α_k assigned to $h_k(\bullet)$.

$$\alpha_k = 0.5 \ln(1/q_k - 1) \tag{7}$$

where $q_t = p_{D_k}[h_k(\mathbf{x}_i) \neq y_i]$

- c. Update distribution $D_{k+1}(i)$ associated with training set

$$D_{k+1}(i) = D_k(i) \exp(-\alpha_k y_i h_k(\mathbf{x}_i)) / Z_k \tag{8}$$

where Z_k is a normalization factor.

3. Output the final hypotheses

$$H(\mathbf{x}) = \sum_{k=1}^T \alpha_k h_k(\mathbf{x}) \tag{9}$$

Our multi-view face detector consists in two level pyramid and six cascades, similar to [8,11], responsible for five different views: frontal view, left-half and left profiles, right-half and right profiles. The first level concerns a cascaded face detector trained on all training examples, which contains non-face examples and face examples in five different views. Only the test image regions which pass the first level will continue to try to pass the second level. Five different cascaded detectors are in the second level which are responsible for detections of faces which may be in different views. There are a total of six cascaded classifiers need to be trained.

A Likelihood Function Based on the Boosted Multi-view Face Detector.

A cascaded detector implicitly assumes a certain form for the underlying probability distribution [12]. Define N_s the total number of layers in the detector, and $1, \dots, n_s$ the layers the detection process passed, in which the output is above the relevant threshold for the input from the test rectangle corresponding to the particle. In our implementation, we assume for simplicity that the likelihood of the particle is related to n_s/N_s . More precisely, we define the structure likelihood as

$$p_s(\mathbf{Y}_k | \mathbf{X}_k) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp - \frac{1 - n_s/N_s}{2\sigma_s^2} \tag{10}$$

where \mathbf{X}_k and \mathbf{Y}_k denote respectively the system state and \mathbf{Y}_k measurement at time step k , σ_s is the standard deviation.

The face detection is performed within the circumscribed rectangle of the minimal area for each particle. Because of the pyramid and the cascade structure, the fast computation of features used in the detector, and the search being constrained in a small image rectangle, the evaluation of the likelihood is efficient. Furthermore, when the test regions contains no face, they will generally fail to pass the first level. This further reduces the computational load.

3 Face Tracking Algorithm Based on Particle Filter

Target tracking can be characterized as the problem of estimating the state \mathbf{X}_k of a system at (discrete) time k , as a set of observations \mathbf{Y}_k become available over time. The Bayesian filtering framework is based on the densities $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ and $p(\mathbf{Y}_k|\mathbf{X}_k)$. The transition prior $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ indicates that the evolution of the state is a Markov process, and $p(\mathbf{Y}_k|\mathbf{X}_k)$ denotes the observation density (likelihood function) in the dynamical system, in which the measurements are conditionally independent of each other given the states. The aim is to estimate recursively in time the posterior density as follows

$$p(\mathbf{X}_k|\mathbf{Y}_{1:k}) = \frac{p(\mathbf{Y}_k|\mathbf{X}_k)p(\mathbf{X}_k|\mathbf{Y}_{1:k-1})}{p(\mathbf{Y}_k|\mathbf{Y}_{1:k-1})} \quad (11)$$

where $\mathbf{Y}_{1:k}$ denotes measurements from the beginning to the current time step k , the prediction density $p(\mathbf{X}_k|\mathbf{Y}_{1:k-1})$ is

$$p(\mathbf{X}_k|\mathbf{Y}_{1:k-1}) = \int p(\mathbf{X}_k|\mathbf{X}_{k-1})p(\mathbf{X}_{k-1}|\mathbf{Y}_{1:k-1})d\mathbf{X}_{k-1} \quad (12)$$

Eq. (11) provides a Bayesian optimal solution of the tracking problem, which, unfortunately, involves high-dimensional integration. In most cases involving non-Gaussianity and nonlinearity, analytical solutions do not exist, leading to the use of Monte Carlo methods.

3.1 Tracking Algorithm Based on Particle Filter

The basic principle of particle filtering [7] is Monte Carlo simulation, that is, the posterior density is approximated by a set of discrete samples (called particles) with associated weights. For each discrete time step, particle filtering generally involves three steps for sampling and weighting the particles, plus one output step. In the sampling step, particles are drawn from the transition prior. In the weighting step, particle weights are set equal to the measurement likelihood. The outputs of the filter are the particle states and weights, used as an approximation to the probability density in state space. In the last step, particles are re-sampled, to obtain a uniform weight distribution. The detailed algorithm is presented as below.

1. Initialisation

Perform exhaustive search with the boosted multi-view detector to detect face. Assume the prior $p(\mathbf{X}_0)$ is Gaussian, from which draw a set $\{(\tilde{\mathbf{X}}_0^{(i)}, 1/N_w), i = 1, \dots, N_w\}$ of particles

2. Sampling step:

For $i = 1, \dots, N_w$: Sample $\mathbf{X}_k^{(i)}$ from the transition prior Eq. (5).

3. Sampling and updating step

- a. Given the particle $\mathbf{X}_k^{(i)}$, evaluate the likelihood $p(\mathbf{Y}_k | \mathbf{X}_k^{(i)})$, as defined in Eq. (10), and set the weight

$$\tilde{w}_k^{(i)} = p(\mathbf{Y}_k | \mathbf{X}_k^{(i)})$$

- b. Normalize the particle weights:

$$w_k^{(i)} = \tilde{w}_k^{(i)} / \sum_{j=1}^{N_w} \tilde{w}_k^{(j)} \quad i = 1, \dots, N_w$$

4. Output step

Output a set $\{(\mathbf{X}_k^{(i)}, w_k^{(i)}), i = 1, \dots, N_w\}$ of particles that can be used to approximate the posterior distribution as $p(\mathbf{X}_k | \mathbf{Y}_{1:k}) \approx \sum_{i=1}^{N_w} w_k^{(i)} \delta(\mathbf{X}_k - \mathbf{X}_k^{(i)})$, and the system mean as the tracking result $\bar{\mathbf{X}}_k \approx \sum_{i=1}^{N_w} w_k^{(i)} \mathbf{X}_k^{(i)}$

5. Selection (resampling) step

Resample the particles $\{(\mathbf{X}_k^{(i)}, w_k^{(i)})\}$ with probability $w_k^{(i)}$ to obtain N i.i.d random particles $\{\tilde{\mathbf{X}}_k^{(i)}, 1/N_w\}$, approximately distributed according to $p(\mathbf{X}_k | \mathbf{Y}_{1:k})$

6. $k = k + 1$, go to step 2.

4 Experiments

The algorithm is implemented with Visual C++ 5.0 on a laptop of Pentium IV-2.2GHz CPU with Microsoft XP. The experiment is made with a real image sequence containing about 500 frames, recorded in a typical office environment, in which both the camera and the subject are moving, and the motion of target is agile and large. The sampling rate is 25Hz and the image is of size 256×192 .

Some of tracking results are shown in Fig. 1. From top to bottom, left to right, illustrated are frames 9, 60, 86, 124, 133, 140, 159, 221, 283, 348, 371, 473, in which the solid red curves represent the tracking results (the system mean $\bar{\mathbf{X}}_k$), and the dashed blue curves represent multi-hypotheses (the particles $\mathbf{X}_k^{(i)}$ whose probability are larger than 0.1). The background is cluttered, due to book shelf, computer screens, window blinds, etc. It can be seen that poses variations are significant, including both in-plane rotation and out-of-plane rotation. The algorithm is also robust to temporary occlusions, tested by a piece of white paper and a hand before the subject face. From about frame 310 to 380, the subject

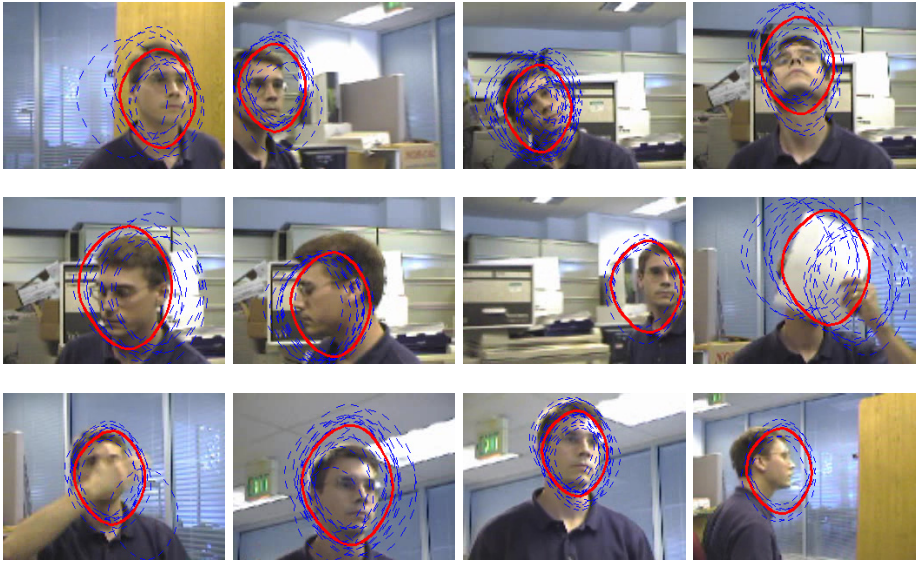


Fig. 1. Some of tracking results in the image sequence. From top to bottom, left to right, illustrated are frames 9, 60, 86, 124, 133, 140, 159, 221, 283, 348, 371, 473, in which the solid red curves represent the tracking results, and the dashed blue curves represent multi-hypotheses. The background is cluttered, due to book shelf, computer screens, window blinds, etc. It can be seen that poses variations are significant, including both in-plane rotation and out-of-plane rotation. The algorithm is also robust to temporary occlusions, tested by a piece of white paper and a hand before the subject face. From about frame 310 to 380, the subject stands up and the illumination changes are strong owing to the lights on the floor. Despite these challenging situations, the algorithm with 20 particles can track the subject face robustly and in real-time.

stands up and the illumination changes are strong owing to the lights on the floor. Despite these challenging situations, the algorithm can track the subject face robustly with $N_w = 20$ particles throughout the whole image sequence, and the mean tracking time for each frame is about 35ms.

We also implement the well-known mean shift algorithm [5] to track faces in the same image sequence. The algorithm will fail when the piece of white paper begins to occlude the face. For the purpose of comparison, we re-initialize the tracker, and it gradually diverges when the subject begins to stand up at about frame 310, because the illumination starts to change gradually but significantly. We herein omit the results of mean shift algorithm for the limit of paper pages. The boosted multi-view face detector itself, or combined with Kalman filter, can be used to track of face, if we perform a local detection on the neighborhood of the result of the previous frame. Unfortunately it will fail in most cases, and can not function as a good tracker.

5 Conclusions

The paper presents a robust and real-time face tracking algorithm. A novel likelihood function is developed based on a boosted multi-view face detector, which is capable of capturing the structural property of the human faces. Furthermore, the likelihood function is integrated with particle filter for robust tracking. The experiments with a challenging image sequence show that the algorithm deal with pose variations of faces, robust to temporary occlusions, illumination changes and cluttered background.

The proposed algorithm is focused on human faces tracking. It is, however, promising and interesting to extend it to tracking of other particular classes of objects, such as pedestrians and cars, the core of which will consist in building likelihoods based on boosted detectors for pedestrians and cars. Future research will focus on these.

References

1. Spors, S., Rabenstein, R.: A Real-time Face Tracker of Color Video. IEEE Int. Conf. on Acoustics, Speech and Signal processing. Utah USA (2001) 1–4
2. Birchfield, S.: Elliptical Head Tracking Using Intensity Gradients and Color Histograms. IEEE Int. Conf. on Comp. Vis. and Pat. Rec. Santa Barbara USA (1998) 232–237
3. Bradski, G.R.: Computer Vision Face Tracking for Use in a Perceptual User Interface. Intel Technology Journal **2**(2)(1998)12–21.
4. Comaniciu, D., Ramesh, V.: Robust Detection and Tracking of Human Faces with an Active Camera. IEEE Int. Workshop on Visual surveillance. Dublin Ireland (2000) 11–18.
5. Comaniciu, D., Ramesh, V., Meer, P.: Real-time Tracking of Non-rigid Objects Using Mean Shift. IEEE Int. Conf. on Comp. Vis. and Pat. Rec. South Carolina USA (2000) 142–149
6. Isard, M., Blake, A.: Cotentour Tracking By Stochastic Propagation of Conditional Density. Proc. European Conf. Comp. Vis. Cambridge UK (1996) 343–356
7. Li, P., Zhang, T., Pece, A.E.C.: Visual Contour Tracking based on Particle Filters. Image and Vision Computing **21** (2003) 111–123
8. Viola, P., Jones, M.J.: Robust Real-time Oject Detection. Workshop on Statistical and Computational Theories of Vision. Vancouver Canada (2001) 26–33.
9. Blake, A., Isard, M.: Active Contours. Springer-verlag, Berlin Germany (1998)
10. Freund, Y., Schapire, R.E.: A Decision-threoretic Generalization of Online Learning and Application to Boosting. J. of Comp. and Sys. Sci. **55**(1) (1997) 119-139
11. Li, S.Z., Xu, L., Zhang, Z., Blake, A., Zhang, H., Shum,H.: Statistical Learning of Multi-view Face Detection. Proc. European Conf. Comp. Vis. Denmark (2002).
12. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: a Statistical View of Boosting. Annals of Statistics. **28** (2000) 337–374

Face Samples Re-lighting for Detection Based on the Harmonic Images

Jie Chen¹, Yuemin Li², Laiyun Qing³, Baocai Yin², and Wen Gao^{1,3}

¹ School of Computer Science and Technology, Harbin Institute of Technology,
Harbin, 150001, China

{chenjie, ymli, lyqing}@jdl.ac.cn

² Multimedia and Intelligent Software Technology Laboratory,
Beijing University of Technology, Beijing 100022, China

ybc@bjut.edu.cn

³ ICT-ISVISION JDL for AFR, Institute of Computing Technology, CAS,
Beijing, 100080, China

wgao@ict.ac.cn

Abstract. Different environment illumination has a great impact on face detection. In this paper, we present a solution by the face relighting based on the harmonic images. The basic idea is that there exist nine harmonic images which can be derived from a 3D model of a face, and by which we can estimate the illumination coefficient of any face samples. To detect faces under the certain lighting conditions, we relight those original face samples to get more new face samples under the various possible lighting conditions by an illumination ratio image and then add them to the training set. By train a classifier based on Support Vector Machine (SVM), the experimental results turn out that the relighting subspace is effective during the detection under the diverse lighting conditions. We also use the relighting database to train an AdaBoost-based face detector and test it on the MIT+CMU frontal face test set. The experimental results show that the data collection can be efficiently speeded up by the proposed methods.

1 Introduction

Over the past ten years, face detection has been thoroughly studied in computer vision research for its interesting applications. Face detection is to determine whether there are any faces within a given image, and return the location and extent of each face in the image if one or more faces present [20]. A hierarchical template matching method for face detection was proposed by Miao et al. [10]. Recently, the emphasis has been laid on data-driven learning-based techniques [8, 9, 14, 15, 19, 21]. However, different environment illumination has a great effect on face detection [20]. To build a robust and efficient face detection system, lighting variations is one of the main technical challenges.

In order to deal with the illumination variations on the faces, many methods have been exploited [1, 7, 11, 13, 16]. Recently, Basri et al. [1] and Ramamoorthi

et al. [13] presented that the appearance of a convex Lambertian object can be well represented with a 9-D linear subspace. By using spherical harmonics and signal-processing techniques, they have shown that the set of images of a convex Lambertian object obtained under varying lighting conditions can be approximated by a 9-D subspace spanned by nine basis images of the object, called harmonic images, each of which corresponds to an image of the object illuminated under harmonic lights whose distributions are specified in terms of spherical harmonics [1, 13]. This discovery has greatly facilitated the modelling of generic illumination and provides the possibility to solve face recognition problem under varying lighting conditions, especially the outdoor environments. The 9-D subspace defined with harmonic images [1] and Harmonic Exemplars [22] provided the possibility to recognize facial images under the diverse lighting conditions. Based on the face relighting model proposed by Basri and Ramamoorthi, Qing proposed the method to relight faces based on illumination ratio image [12]. Chen presents a genetic algorithm (GA)-based method to swell face database through re-sampling from existing faces and re-lighted the samples by the linear point light source [2].

In this paper, we propose a method for face relighting based on harmonic images [1, 13]. It is to relight face samples under any certain illumination to simulate the possible lighting variations of faces in the images. By this scheme, we can enrich the lighting variations of the training set. Using these newly produced samples together with the original, we train a classifier SVM and prove that the hit rates can be improved by this method.

The rest of this paper is organized as following: In section 2, we introduce the method to relight face samples to the certain illumination based on the harmonic images. The experiment results are described in Section 3. In Section 4, we give the conclusions.

2 Face Relighting with the Spherical Harmonics

As discussed in [1, 13], assuming a face is a convex Lambertian surface, we can denote the face image:

$$I(x, y) = \rho(x, y) \vec{n}(x, y) \vec{s} \quad (1)$$

where $\rho(x, y)$ is the albedo of the point (x, y) ; $\vec{n}(x, y)$ is the surface normal direction; \vec{s} is the point light source direction, whose magnitude is the light source intensity.

In space-frequency domain, Lambertian surface is a low-pass filter and the set of images of a Lambertian object under varying lighting can be approximated by a 9D linear subspace spanned by the harmonic images b_{lm} ($0 \leq l \leq 2$, $-l \leq m \leq l$)[13]. The harmonic images are defines as:

$$b_{lm}(x, y) = \rho(x, y) A_l Y_{lm}(\theta(x, y), \phi(x, y)), \quad (2)$$

where Y_{lm} is the spherical harmonic at the surface normal; (θ, ϕ) a pair of angles corresponding to the pixels, is the function of (x, y) and $0 \leq \theta \leq \pi$, $0 \leq \phi \leq 2\pi$; $A_l (A_0 = \pi, A_1 = 2\pi/3, A_2 = \pi/4)$ is the spherical harmonics coefficients.

The image under the arbitrary lighting can be written as:

$$I(x, y) = \sum_{l=0}^2 \sum_{m=-l}^l L_{lm} b_{lm}, \quad (3)$$

where L_{lm} is the spherical harmonic coefficients of the specific lighting. The nine lower spherical harmonic coefficients L_{lm} ($0 \leq l \leq 2$) can be estimated as discussed in [13]. Given an input image \mathbf{I} (a column vector of n elements, n is the number of pixels in an image), if the harmonic images of an object are known, then the coefficients of the illumination \mathbf{L} can be solved by the least squares problem:

$$\hat{\mathbf{L}} = \arg \min \|\mathbf{B}\mathbf{L} - \mathbf{I}\|, \quad (4)$$

where \mathbf{B} denotes the harmonic images, arranged as a $n \times 9$ matrix. Every column of \mathbf{B} contains one harmonic image b_{lm} as in Eq. (2).

Assumes a convex Lambertian object in a distant isotropic illumination, its irradiance has been proved that its most energy is constrained in the three low order frequency components and its frequency form can be formulated as [1] [13]:

$$E(\theta, \phi) \approx \sum_{l=0}^2 \sum_{m=-l}^l A_l L_{lm} Y_{lm}(\theta, \phi). \quad (5)$$

Once we have estimated the lighting of the original image as in Eq.(4), it is commonsense to relight it to the canonical illumination with the illumination ratio image. According to Eq. (2), (3), (5), illumination ratio image between the canonical image and the original image is defined as [12]:

$$IRI(x, y) = \frac{I_{can}(x, y)}{I_{org}(x, y)} = \frac{E_{can}(\theta(x, y), \phi(x, y))}{E_{org}(\theta(x, y), \phi(x, y))}, \quad (6)$$

where the subscripts are the illumination indexes, and E is the incident irradiance. Relighting image with the illumination ratio image can be rewritten as:

$$I_{can}(x, y) = IRI(x, y) \times I_{org}(x, y). \quad (7)$$

The ratio-image above defined is almost useless since it is only applicable to the original face. However, noticing that all faces have similar 2D and 3D shapes, we can firstly warp all faces to the generic shape and then compute the ratio-image for relighting the face images. It is then easy to reversely warp the relighting image back to its original shape. Currently, we just warp the 2D face image to a predefined mean shape, as defined in ASM. After the warp procedure, all face images are expected to have quite similar 3D shape. Using the face relighting method, we can get samples under the specific illumination conditions. Some relighting examples are illustrated in Fig. 1.

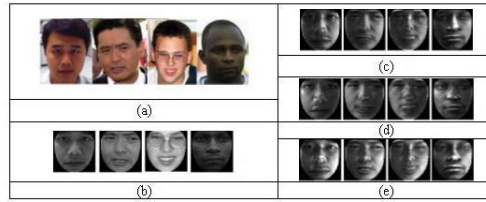


Fig. 1. Face samples relighting. (a) The original image; (b) cropped, normalized and masked face samples; (c) (d) (e) relighting under different lighting conditions.

3 Experiment Results

3.1 Face-Samples Preprocessing

The data set is consisted of a training set of 6977 images (2429 faces and 4548 non-faces) and two test sets. The first test set (Set1) is composed of 24045 images (472 faces and 23573 non-faces). All of these images were 19×19 grayscale and they are available on the CBCL webpage [24]. The second test set (Set2) is a subset of CAS-PEAL, which can be downloaded from [5]. And the subset includes: CAS-PEAL_Up_0, CAS-PEAL_Dn_0, CAS-PEAL_Mid_0, where CAS-PEAL_Up_0 means those images are illuminated by the 0 degree light source form overhead, CAS-PEAL_Dn_0 and CAS-PEAL_Mid_0 are also be captured under 0 degree light source. Some examples are shown in Fig.2, where the arrows denote the direction of the light source. Using the method presented above, we can relight those face samples in the training set.

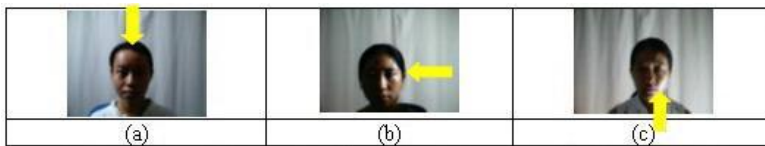


Fig. 2. Some test samples of CAS-PEAL. (a) is an example from CAS-PEAL_Up_0; (b) is a example from CAS-PEAL_Mid_0; (c) is a example from CAS-PEAL_Dn_0.

3.2 Comparing the Solutions Performance

To train and test the Support Vector Machines (SVMs) [18], we use the SVMFu version 2.001[23]. We train SVMs using the grayscale values and a polynomial kernel of degree 2. Fig. 3 provides the results for the classifier SVMs trained with different database and tested on Set1. In this figure, we use only the initial data set of CBCL (No-Light), and the initial database together with others 1000 face

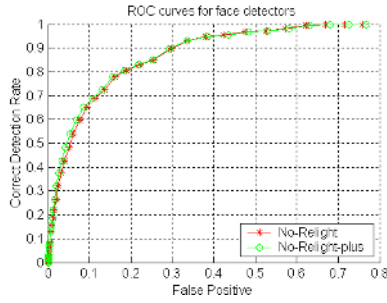


Fig. 3. The ROC curves by adding training face samples and being tested on Set1.

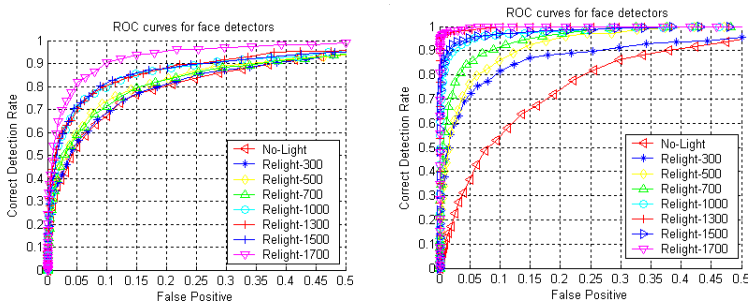


Fig. 4. The ROC curves of the trained SVM tested on the test sets. (a) The detect results on Set1; (b) the detect results on Set2.

samples collected from webpage (called No-Light-plus here). It means the No-Light has 2429 face samples, while No-Light-plus has 3429 face samples. The test set is the Set1 as discussed in Section 3.1. From these Receiver Operating Characteristic (ROC) curves in Fig. 3, we can find that the performance improvement of No-Light-plus is much limited compared with that of the No-Light. That is to say only expanding the number of the training set without considering the lighting variations can not improve the performance of the detector distinctly.

Fig. 4 provides the results for the classifier SVMs trained with others different database and tested on the testing sets. In these figures, we use only the initial data set of No-Light, and part of the initial database together with others different number face samples, which have been relighted by the methods demonstrated in section 2. Herein, No-Light is those face samples of CBCL database; Relight-300 means we substitute 300 relighting face samples for the same number of samples of CBCL database; and the same is of Relight-500, 700, ..., 1700. Note all of these eight cases have the same face samples (that is to say 2429 positive examples.). The trained classifier SVMs, by these eight different positive sample sets and the same negative samples of the CBCL database, are tested on Set1 and Set2. The results are illustrated in Fig. 4. One can find that

the presented method can improve the performance of the classifier. Comparing with the former results illustrated in Fig. 3, we can conclude that the relighting samples are more useful to train the face detectors under varying lighting conditions than simply adding the number of the training samples. However, the improvement of the performance on Set2 is more distinct than that of on Set1. It is because more lighting variations of Set2 contribute to it. It also demonstrates that this scheme will be more efficient for those images with diverse lighting conditions, which is just the problem of other face detectors. From these ROC curves, one can conclude that the performance improvement will decrease with the increasing of the substituted samples, for example, from No-light to Relight-300 compared with from Relight-1500 to Relight-1700. It may be that the certain number of lighting samples can represent the possible lighting variations and it makes this scheme more practical. We get the same results by using different original samples to do these experiments. That means the results are only related to the lighting conditions we relight the samples while it has little relations to the samples themselves.

3.3 Evaluation of the Generated Samples

Considering that the solutions performance of the relighting samples is evaluated by the classifier SVM above, they may favor this classifier. In order to verify that the solutions are independent to any special classifier, we use the relighting training set to train another classifier and test its generalization performance. AdaBoost has been used in face detection and is capable of processing images extremely rapidly while achieving high detection rates [19]. Therefore, we use the AdaBoost algorithm to train a classifier. A final strong classifier is formed by combining a number of weak classifiers. For the details of the AdaBoost based classifier, please refer to [3].

To compare the performance improvement on different training sets, we also use two different face training sets. The face-image database consists of 6,000 faces (collected from Web), which cover wide variations in poses, facial expressions and also in lighting conditions. To make the detection method less sensitive to affine transform, the images are often rotated, translated and scaled [4, 6, 8, 9, 14, 21]. Therefore, we randomly rotate these samples up to $\pm 15^\circ$, translate up to half a pixel, and scale up to $\pm 10\%$. We get 12,000 face images. And these 12,000 face images constitute the first group *No-Relighting*. The second group *Relighting* is composed of 10,000 face images as above-mentioned and 2,000 relighting samples by the proposed method.

The non-face class is initially represented by 10,000 non-face images. Each single classifier is then trained using a bootstrap approach similar to that described in [17] to increase the number of negative examples in the non-face set. The bootstrap is carried out several times on a set of 12,438 images containing no faces.

The resulting detectors, trained on the two different sets, are evaluated on the MIT+CMU frontal face test set which consists of 130 images showing 507 upright faces [14]. The detection performances on this set are compared in Fig. 5.

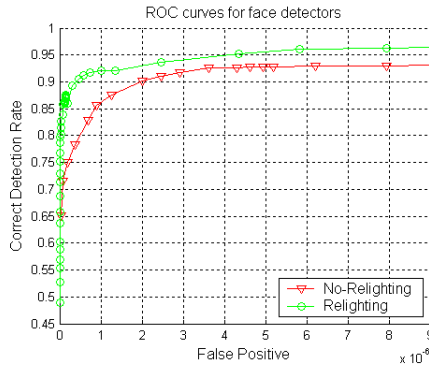


Fig. 5. The ROC curves of the AdaBoost-based classifiers are tested on test sets.

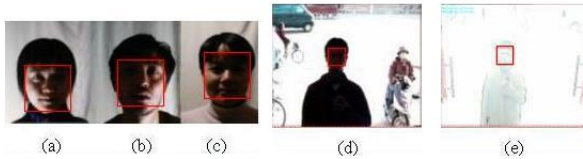


Fig. 6. Some face detection results; (a), (b), (c) are detected and cropped from CAS-PEAL database; (d), (e), (f), (g) from the practical applications.

From the ROC curves one can find that we get the detection rate of 90.8% and 15 false alarms with the detector trained on the set *Relighting*. Viola reported a similar detection capability of 89.7% with 31 false detects (by voting) [19]. However, different criteria can be used to favor one over another, which will make it difficult to evaluate the performance of different methods even though they use the same benchmark data sets [20]. Some results of this detector are shown in Fig. 6.

4 Conclusions

In this paper, we present a novel method to relight face training set. This scheme can improve the hit rates of the classifier under the diverse lighting conditions. The experiment results based on SVM and AdaBoost-based classifiers demonstrate the generation performance and its independence on the specific classifier.

Acknowledgements. This research is partially sponsored by Natural Science Foundation of China under contract No.60332010 and No.D070601-01, National Hi-Tech Program of China (No. 2001CCA03300, 2001AA114160, 2001AA114190 and 2002AA118010), and ISVISION Technologies Co. Ltd.

References

1. R. Basri and David W. Jacobs. Lambertian Reflection and Linear subspaces. *IEEE transactions on Pattern Analysis and Machine Intelligence* Vol. 25, No. 2. Feb. 2003.
2. J. Chen, X. L. Chen, W. Gao. Expand Training Set for Face Detection by GA Re-sampling. *FG2004*. pp. 73-79.
3. Y. Freund, R. E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Computational Learning Theory*, 1995. pp. 23-37.
4. B. Froba and A. Ernst. Fast Frontal-View Face Detection Using a Multi-Path Decision Tree. *Proc. AVBPA2003*, pp. 921-928.
5. W. Gao, B. Cao, S. G. Shan, D. L. Zhou, X. H. Zhang, D. B. Zhao. The CAS-PEAL Large-Scale Chinese Face Database and Evaluation Protocols. Technical Report. 2004. (<http://www.jdl.ac.cn/peal/JDL-TR-04-FR-001.pdf>)
6. B. Heisele, T. Poggio, and M. Pontil. Face Detection in Still Gray Images. *CBCL Paper #187*. Massachusetts Institute of Technology, Cambridge, MA, 2000.
7. K. C. Lee, J. Ho, D. Kriegman. Nine Points of Lights: Acquiring Subspaces for Face Recognition under Variable Illumination. *CVPR2001*, pp. 519-526.
8. S. Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H. J. Zhang, and H. Shum. Statistical Learning of Multi-View Face Detection. In *Proceedings of the 7th ECCV*. 2002.
9. C. J. Liu. A Bayesian Discriminating Features Method for Face Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, June 2003. pp. 725-740.
10. J. Miao, B.C. Yin, K.Q. Wang, et al., A Hierarchical Multiscale and Multiangle System for Human Face Detection in a Complex Background Using Gravity-Center Template, *Pattern Recognition* 32(7), 1237-1248, 1999
11. P. J. Phillips, P. Grother, R.J. Micheals, et. al.. *FRVT 2002: Evaluation Report*, http://www.frvt.org/DLs/FRVT_2002_Evaluation_Report.pdf, March 2003.
12. L. Y. Qing, S. G. Shan, X. L. Chen. Face Relighting for face recognition under generic illumination. *ICASSP2004* (Accepted).
13. R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. *Journal of the Optical Society of America*. Vol.18, no.10, pp.2448-2459, 2001.
14. H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *IEEE Tr. Pattern Analysis and Machine Intel*, vol. 20, 1998, pp. 23-38.
15. H. Schneiderman and T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces. *Computer Vision and Pattern Recognition*, 2000. pp. 746-751.
16. A. Shashua and T. Riklin-Raviv. The Quotient Image: Class-Based Re-Rendering and Recognition With Varying Illuminations. *IEEE TPAMI*, pp.129-139, 2001.
17. K. K. Sung, and T. Poggio. Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. on PAMI* Vol.20. , No. 1, 1998. pp. 39-51.
18. V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
19. P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. *Conf. Computer Vision and Pattern Recognition*, 2001. pp. 511-518.
20. M. H. Yang, D. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Tr. Pattern Analysis and Machine Intelligence*, vol. 24, Jan. 2002. pp. 34-58.
21. M. H. Yang, D. Roth, and N. Ahuja. A SNoW-Based Face Detector. *Advances in Neural Information Processing Systems 12*, MIT Press, 2000. pp. 855-861.
22. L. Zhang, D. Samaras. Face Recognition under Variable Lighting using Harmonic Image Exemplars. *Proc. CVPR'2003*, Vol. I, pp. 19-25, 2003.
23. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
24. <http://www.ai.mit.edu/projects/cbcl/software-dataset/index.html>.

A Dictionary Registration Method for Reducing Lighting Fluctuations in Subspace Face Recognition

Kenji Matsuo, Masayuki Hashimoto, and Atsushi Koike

KDDI R&D Laboratories Inc., 2-1-15 Ohara, Kamifukuoka-shi,
Saitama 356-8502, Japan

Abstract. A dictionary registration process of subspace face recognition for realizing strong tolerances to lighting fluctuations is described in this paper. Camera devices have been added to handy phones and enable us to do biometric certification using face images. However, face images taken by handy phones are easily fluctuated by lighting conditions. Our proposed method is based on subspace face recognition which can realize robustness against fluctuations. Proposed method can create a subspace with robustness against lighting fluctuations virtually, using the lighting canonical space which contains various types of lighting elements. Not only equal-error-rate of verification but also precision of identification can be improved by proposed method, from 48.7% up to 91.9%.

1 Introduction

Various services are expanded on handy phone, such as mobile banking, stock trade and on-line services. Handy phone holders have desired a new certification scheme with high security. Meanwhile biometric certification can realize the highest security. Therefore, biometric certification using face images taken by handy phone have been developed. Recently, camera devices have been added to almost all handy phone. It has made biometric certification of face images possible without the addition of extra devices, such as a fingerprint sensor. There are already many elementary techniques in the face recognition field[1], for example, eigenface[2], LFA[3], gabor wavelet[4] and so forth. Especially, the subspace method[5] can be applied to face recognition on handy phones, even if handy phone resources are severely restricted. This method can handle a low resolution face image and projects it into a subspace which is represented by a lower dimension than original. However, the face images taken by handy phones are easily fluctuated by lighting conditions. They also decrease the certification precision, because handy phone are used both indoors and outdoors.

In this paper, we propose a dictionary registration process of subspace face recognition for realizing strong tolerances to lighting fluctuations. This method uses the lighting canonical space which contains various types of lighting elements and was modeled in advance and creates a subspace virtually without actually taking face images under various conditions. This paper is composed as follows. In Chapter 2, face recognition using the conventional subspace method

and its problem are described. In Chapter 3, how to create the subspace virtually is proposed. In Chapter 4, simulations are performed and advantages of the proposed method are confirmed. This paper is concluded in Chapter 5.

2 Subspace Face Recognition

The subspace method[5] has been a popular method for pattern recognition and one of its features is the dimension reducing effect. When N face images $\mathbf{x}_i (i = 1, 2, \dots, N)$ of a certain person are obtained, the correlation matrix R of these face images is defined as the following equation:

$$R = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T \quad (1)$$

Vectors $\mathbf{e}_j (j = 1, 2, \dots, N)$ are the eigen vectors of R and are also arranged in the same order corresponding to each eigen value. M eigen vectors become the orthogonal base spanning the face image subspace. M is less equal than N and M eigen vectors of all are selected in decreasing order. This method judges who the person is, measuring the similar S or angle θ between the input face image \mathbf{x} and the subspace spanned by the M eigen vectors, defined as the following equation:

$$S = \cos^2 \theta = \sum_{j=1}^M \frac{(\mathbf{e}_j^T \cdot \mathbf{x})^2}{|\mathbf{e}_j|^2 |\mathbf{x}|^2} \quad (2)$$

If the person of the input face image \mathbf{x} is the same as his own subspace, S shows a large value near 1. If not, S shows a small value near 0. That is to say, judgment of recognition can be determined by measuring whether S is close to 1 or not.

The conventional subspace method has one serious problem: the subspace method can work effectively only when the registered subspace contains various types of lighting fluctuation elements. That is, if the registered subspace contains only one or a few types of lighting fluctuation elements, its subspace cannot be perfectly adapted to various types of lighting fluctuations.

Assuming that face image \mathbf{x} was fluctuated by a different lighting condition from that of the dictionary registration process and is used for the recognition process, similar S between the input face image \mathbf{x} and subspace becomes a small value and also recognition accuracy could decrease. To provide strong tolerances to lighting fluctuation for the subspace, many face images under various lighting conditions must be taken in the registration process. A subspace containing various types of lighting fluctuation elements are created from these face images, the recognition process prevents similar from decreasing and consequently solves the problems on lighting fluctuations.

3 Proposed Registration Process

The conventional subspace method can work against lighting fluctuations only if the subspace is made from many face images and these face images have been

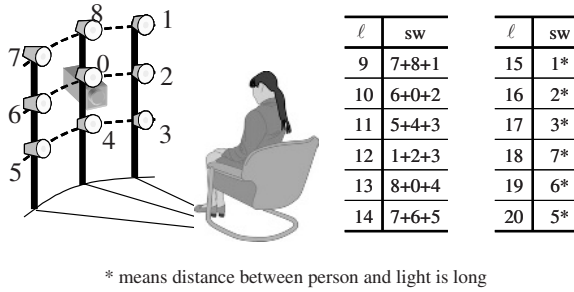


Fig. 1. Experimental environment and lighting conditions ℓ

taken under various types of lighting conditions. However, it is actually difficult to perform this operation, since the number of face images taken should be reduced so that the registration process of this system can be used as simply and as easily as possible. In this situation, the subspace contains only one or a few types of lighting fluctuation elements and subspace face recognition doesn't have tolerances to the change in the lighting condition. In this Chapter, we propose a new registration method of the subspace dictionary that can recognize fluctuated face images correctly even in such a situation. The lighting canonical space is modeled from many face images which have been taken under various types of lighting condition in advance. Using this lighting canonical space, the proposed subspace with various types of lighting fluctuation elements is virtually reconstructed from the conventional subspace.

Experimental environment and lighting conditions used in Chapter 3 and Chapter 4 are shown in Figure 1.

3.1 Lighting Canonical Space

Face images $\mathbf{y}_{p\ell}$ of P persons ($p = 1, 2, \dots, P$) are taken under L types of lighting conditions ($\ell = 1, 2, \dots, L$), respectively. From this set of PL face images, its eigen space can be introduced in advance and various types of lighting fluctuation elements can also be modeled into this eigen space. This modeled space is called the lighting canonical space and this face image set is also called the canonical face image set. Mean $\bar{\mathbf{y}}$, center vector \mathbf{m} and covariance matrix C of the lighting canonical space are defined as the following equations:

$$C = \frac{1}{PL} \sum_{p=1}^P \sum_{\ell=1}^L (\mathbf{y}_{p\ell} - \mathbf{m})(\mathbf{y}_{p\ell} - \mathbf{m}) \tag{3}$$

$$\mathbf{m} = \bar{\mathbf{y}} = \frac{1}{PL} \sum_{p=1}^P \sum_{\ell=1}^L \mathbf{y}_{p\ell} \tag{4}$$

$PL - 1$ eigen vectors can be introduced from the principal component analysis of C , and they become the orthogonal base[1] spanning the lighting canonical space.

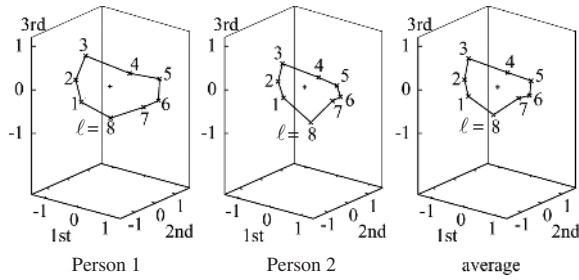


Fig. 2. Manifold curves on lighting canonical space

Next, face images are taken as lighting conditions are changed continuously. Then these face images are mapped onto the lighting canonical space, as shown in Figure 2, where only three of the foremost eigen vectors were selected in decreasing order of eigen value and corresponding projection vectors were plotted. This graph shows the following two properties[6]. 1) The track of their projection vectors depicts a continuous manifold curve and these curves are not dependent on persons but are an almost similar shape. 2) In the lighting canonical space, only a few of the foremost eigen vectors represent the majority of the elements of lighting fluctuations. Therefore, the proposed method is based on the above two properties. This method computes the average curve in advance. Maintaining the same position relation of the projection vectors as this average curve, another face image can be deduced from a certain face image by operating only a few of the foremost coefficients of its projection vector.

Figure 3 shows the virtual face images under 8 types of lighting conditions, in 8 directions such as up, down, left, right and sideways. It is confirmed that the virtual face images under other lighting conditions can be made by the above operation.

3.2 Creation Algorithm of Virtual Image and Virtual Subspace

L virtual face images could be deduced respectively from one face image using the continuous curve on the lighting canonical space. If N face images \mathbf{x}_i ($i = 1, 2, \dots, N$) were taken in the registration operation, NL virtual face images can be obtained as a consequence. These virtual face images can create a subspace which contains L types of lighting fluctuation elements. Its correlation matrix R' can be introduced from equation (1). The virtual subspace introduced from R' has strong tolerances to lighting fluctuations, since R' contains L types of lighting fluctuation elements. From the above process, we can achieve a face image recognition that has two special features, that is, 1) strong tolerances to lighting fluctuations and 2) high usability without actually taking face images under various lighting conditions.

Moreover, R' can be introduced from the summation of the two matrixes R and C simply, only if the lighting condition in the center of the lighting canonical

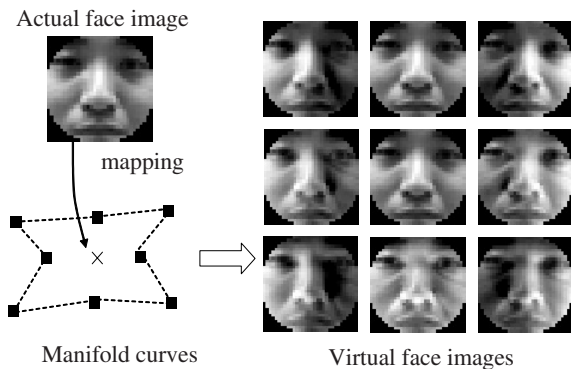


Fig. 3. Deducing operation on lighting canonical space

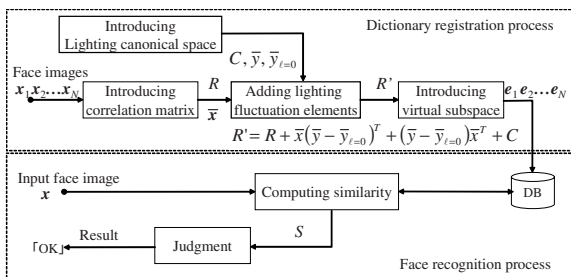


Fig. 4. Flow chart of proposed face recognition

space is consistent with the lighting condition in the registration process.

$$R' = R + C \tag{5}$$

However, such consistency is very rare and the lighting condition in the center of the lighting canonical space is usually different from the lighting condition in the registration process. Therefore, to realize a user-friendly face recognition system, this inconsistency must be considered here and we propose the following face registration method and its face recognition system. Now, the lighting condition in the registration process is at any time fixed to a certain lighting condition, that is defined as $\ell = 0$. In advance, the face images for the lighting canonical space are also taken under $L + 1$ types of lighting conditions including $\ell = 0$. \bar{x} is the mean of the face images, which were taken under the same lighting condition $\ell = 0$. $\bar{y}_{\ell=0}$ is used as the center vector m of the lighting canonical space instead of \bar{y} . R' can be introduced using R , \bar{x} , $\bar{y}_{\ell=0}$, \bar{y} , and C , as shown in the following equation.

$$R' = R + \bar{x}(\bar{y} - \bar{y}_{\ell=0})^T + (\bar{y} - \bar{y}_{\ell=0})\bar{x}^T + C \tag{6}$$

Figure 4 shows the flow chart of the proposed dictionary registration algorithm and the face recognition algorithm.

Table 1. Classification of test face images

Lighting condition ℓ	(A)10 persons	(B)11 persons		
	for C matrix	for R matrix	for R'' matrix	for experiment
0	5	50		5
1 ~ 8	each 5	0	each 5	each 5
9 ~ 20	0	0	0	each 5
Total	450	50/person	90/person	1155

4 Computer Simulations

Computer simulations were performed to check the advantages of the proposed method. Face images taken in our laboratories were used in these simulations. This dataset is composed of 21 persons and the face images of each person were taken under 21 types of lighting condition. The resolution of all face images was normalized so that the width was 32 and the height was 32. All 21 persons were classified into two groups. The first group (A) was for making the covariance matrix of the lighting canonical space. The second group (B) was for measuring recognition accuracy. This classification of test face images is shown in Table 1. First, for $P = 10$ persons in group (A), C of the lighting canonical space was introduced from the face images, which were taken under $\ell = 0$ and $L = 8$ types of lighting conditions, containing $\ell = 1 \sim 8$. Next, for 11 persons in group (B), which did not contain the same persons in group (A), the correlation matrix of subspace R was introduced from $N = 50$ face images, which were taken under the $\ell = 0$ lighting condition only.

4.1 Effects on Verification

1:1 verification precision of the face images fluctuated by changes in the lighting condition was measured in about 11 persons in group (B). Each R' was introduced simply from the addition of R and C as shown in equation (6), and each subspace was created from this R' . 8 types of lighting conditions, that is $\ell = 1 \sim 8$, were contained in both groups (A) and (B). However, the other 12 types of lighting conditions, that is $\ell = 9 \sim 20$, were not contained in the lighting canonical space of group (A) and were contained only in group (B). Similarities were measured for all 21 types of lighting conditions. The conventional subspace and proposed virtual subspace were set to 50 dimensions, that is, 50 eigen vectors were used in this experiment.

First, the False Accept Rate (FAR) and False Reject Rate (FRR) were calculated with the change in threshold. This result is shown in Figure 5. ERR of the conventional subspace method is about 0.36. In contrast, the proposed subspace method can archive about 0.23. It was confirmed that the proposed method could improve the verification precision and reduce recognition error. In addition, although the lighting canonical space contains only 8 types of lighting conditions, the proposed subspace could also verify the face images that

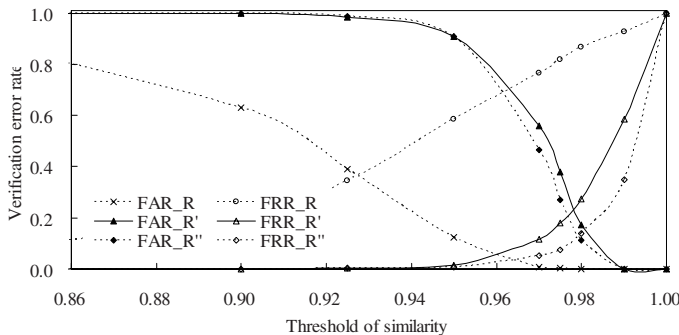


Fig. 5. FAR and FRR of verification

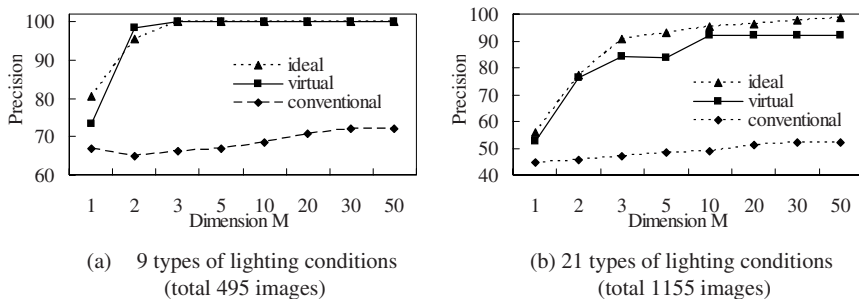


Fig. 6. Identification precision

were taken under other different lighting conditions not contained in the lighting canonical space. This result shows that arbitrary lighting fluctuations can be calculated by the linear summation of representative lighting fluctuation elements which are contained in the lighting canonical space. Here, assuming that all face images of each person belonging to group (B) are ideally taken under $\ell = 0$ and 8 types of lighting conditions $\ell = 1 \sim 8$ in the registration process, ideal correlation matrix R'' can be introduced from their 90 face images. FAR and FRR of the ideal subspace introduced from R'' are also calculated and are shown in Figure 5. ERR of the ideal subspace shows about 0.13. In contrast, the proposed virtual subspace can archive about 0.23. From these results, the virtual subspace introduced from R' has high performance closest to the ideal subspace introduced from R'' .

4.2 Effects on Identification

The proposed subspace method adds C equally to R of anyone and creates R' . Therefore, there are some concerns that the proposed subspace may improve similar in anyone equally and finally. If so, the segmentation capability among classes of virtual subspace is low and its identification precision may be equal to the conventional subspace method. However, such a problem does not in

fact occur, since subordinate components between R and C are cancelled in the process when eigen vectors of R' are introduced and individual features can be correctly reflected in the subspace.

We measured 1:n persons identification precision to verify the justification of this theory. This result is shown in Figure 6. R means the conventional subspace and R' means the proposed subspace. It was confirmed that the proposed method has higher identification precision than the conventional method in both cases. The first case (a) contains 1+8 types of the same lighting conditions, that is $\ell = 0, 1 \sim 8$, in the lighting canonical space. The second case (b) contains 1+20 types of lighting conditions that is $\ell = 0, 1 \sim 20$, which are partly not included in the lighting canonical space. Moreover, the proposed method can improve the identification precision more as the number of used eigen vectors M , that is, as the subspace dimension increases, and can archive 100% in case (a) and 91.9% even in case (b). 1:n persons identification precision of ideal subspace introduced from R'' are also calculated and are shown in Figure 6. The ideal subspace achieves 100% identification precision in case (a) and 95.2% even in case (b). From this results, virtual subspace introduced from R' has high performance closet to ideal subspace introduced from R'' .

5 Conclusion

We propose a dictionary registration process of subspace face recognition for realizing strong tolerances to lighting fluctuations. The proposed method create a subspace virtually using the lighting canonical space. More over, setting a lighting condition in registration process to the same lighting condition of center of canonical space, proposed method can improve faithfulness of viatual subspace. Consequently, proposed method can improve verification error rate from 0.36 to 0.23 and identification precision from 48.7% to 91.9% and this method becomes one of the useful techniques for biometric certification using handy phones.

References

- [1] E. HjeIm and B. K. Low, "Face Detection: A Survey," Computer Vision and Image Understanding, Vol. 83, No. 3, pp. 236-274, Sept. 2001.
- [2] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," Proceedings Of IEEE Conference on Computer Vision and Pattern Recognition, pp. 586-591, Jun. 1991.
- [3] P. S. Penev and J. J. Atick, "Local feature analysis: a general statistical theory for object representation," Network: Comput. Neural Syst. 7, pp. 477-500, 1996.
- [4] J. G. Daugman, "Complete Discrete 2-D Gabor Transform by Neural Networks for Image Analysis and Compression," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 36, No.7, pp. 1169-1179, 1988.
- [5] E. Oja, "Subspace methods of pattern recognition," Research Studies Press, Hertfordshire, 1983.
- [6] H. Murase and S. K. Nayar, "Illumination Planning for object recognition using parametric eigenspace," IEEE Trans. Pattern Anal. & Mach. Intell., Vol. 16, No. 12, pp. 1219-1227, 1994.

A 3D-Dialogue System Between Game Characters to Improve Reality in MMORPG

Dae-Woong Rhee, Il-Seok Won, Hyunjoo Song, Hung Kook Park, Juno Chang,
Kang Ryoung Park, and Yongjoo Cho

Sangmyung University, 7 Hongji-dong, Jongno-gu, Seoul, 110-743, Republic of Korea

Abstract. In the Persistent World of the cyberspace formed by MMORPG, realistic conversations between game characters play an important role so that players can be immersed in the game. However, this conversation between game characters in MMORPG has been limited to simple 2D-Dialogues. In other words, when a PC-the game character controlled by the players-and a NPC-the character provided and controlled by the back-end game system, they only communicate simply based on the levels of PCs and the types of NPCs. In this paper, we attempt to extend this system to 3D-Dialogue System by adding familiarity degree between two characters.

1 Motivation

With the big growth of online game industry in late 1990s, a new era of game development began and it drew great attention from the industry and research communities. Especially, MMORPG (Massively Multi-Player Online Role Playing Game) contributed to bring the games to the public and generated great number of game players.

However, the concept of multi-player online games existed before. The multi-player online games have been developed since mid-1970s and grouped as the multi-user games. Nevertheless, MMORPG evolved from other online multi-user games and opened up a new subarea in the online game genre because it was able to differentiate itself from other online games with a new characteristic of providing Persistent Worlds (hereafter PWs) [1]. The PWs refer to persistent virtual worlds in the cyberspace that is lasted and even evolved by game players.

In MMORPGs, many users have their own characters and play at any time by logging in the virtual worlds and exit the game by logging out. Because the MMORPG's PWs are extensively large and continuous, players' immersion are varied by how realistically the worlds are presented to the users. Thus, a high fidelity of virtual worlds becomes an important design issue in MMORPGs.

Since Nexon made the debut of the world's first MMORPG game, "Kingdom of the Winds" [2] in 1994, the development of MMORPGs showed the remarkable growth with the advancement in the computer technologies, such as 3D graphics technologies, artificial intelligence techniques, high resolution displays, 5.1 channels surround sound technologies, and so on.

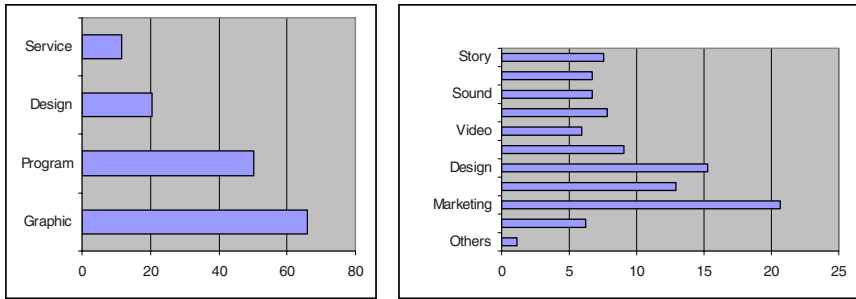


Fig. 1. The left image shows the result of the preference survey of MMORPG service. The right image shows what technology is currently deficient in the online game companies in Korea

However, user satisfaction of MMORPG is not surprisingly high even with the advanced computer technologies. According to the preference survey of MMORPG players among five hundred men and women in Korea, only 23.2 percent of the users (about 116 players) reported that they were satisfied with the current service. As shown in Fig 1, the detailed surveys showed the lowest satisfaction in service and design: only 11.6 percent of the users were satisfied with the online game services and 20.6 percent of the users were satisfied with the overall game design. This result is similar to the report by KGDI (Korean Game Development & Promotion Institute) that studied the deficient technologies or skills in existing online game companies. According to the study, the design was ranked secondly as an urgent need for improvement, followed by the advertisement/marketing (see the right image in Fig. 1). Therefore, both the companies and players indicated that the design of the overall games should be improved in the development of MMORPGs.

The following section describes the definition of each parameter in the 3D-Dialogue System and some examples showing the depth of the conversation caused by the familiarity. Finally, it also describes the applications of 3D-Dialogue System and its future research direction.

2 2D-Dialogue System

The elements of game design consists of several elements, such as the storyline, scenario, game play interface, and rules. In this paper, we studied a method to increase the immersion of the game players by providing the solid dialogues among game characters. Non-playable characters (hereafter NPCs) are autonomous agents that cannot be controlled directly by the players. However, NPCs stay in the game server all the time and help the players get experience of PWs. NPCs play many roles in the games, such as telling the story of the game environments, generating events, supporting buying or selling game items of the virtual worlds. Then, game players control the playable characters (hereafter PCs) and

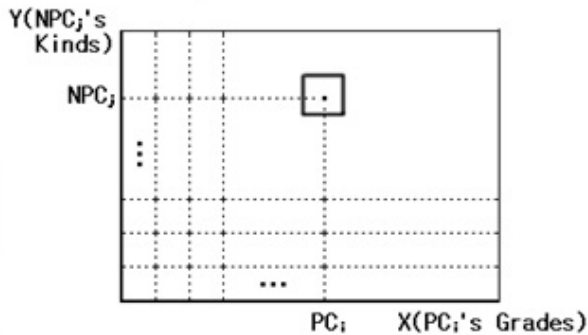


Fig. 2. 2D-Dialogue System

experience the PWs. Thus, the communication between PCs and NPCs plays a vital role in the virtual worlds, and hence enabling more realistic communication between these characters helps players be immersed more into the games.

In the existing MMORPGs, the communication between PC and NPC is done by simple questions and answers, i.e. a simple dialogue where a PC asks a question to a NPC and the NPC answers it. Although some NPCs talk by themselves, it happens in special occasions for specific events. NPCs never initiate talking to the PCs. The answers are often presented as simple choice questions whether the responding NPCs would select simple choices or not.

Nexon's "Kingdom of the Winds" shows the early generation of MMORPG's NPCs. There are only small numbers of NPCs available in the game, and their dialogues are usually greetings and typical conversations. Non-typical conversations are implemented as special events of the game. However, a recent product called Mabinogi [4] developed by Nexon shows a new trend of MMORPG developments.

Mabinogi emphasizes on building a community of the game characters including both PCs and NPCs. The NPCs of the game generated more dynamic expressions and provides several levels to overcome the simple characteristics of the old generation of MMORPGs. Unlike the NPCs of "Kingdom of the Winds", Mabinogi's NPC can show more diverse expressions. For example, if a PC contacts a NPC in Mabinogi, NPC asks a question "how can I help you?" to the PC and provides several options, such as [talk], [sell or buy], [repair an item], which can be chosen by the PC.

Grabity's Ragnarok [5], which is well known as the game that allows building communities of PCs and NPCs, provides a variety of NPCs. It also provides more specialized groups and guild services to support the activities of the communities. Simple dialog box user-interface, improved guilds or battles, introduction of PvP (Player vs. Player) zones, and enhanced PC's power, the items, pet system, increased numbers and types of NPCs, and part-time works are also provided to support various community-oriented game activities. According to the analysis

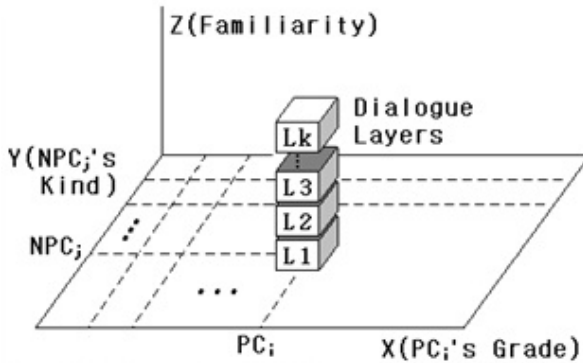


Fig. 3. 3D-Dialogue System

of more than twenty MMORPGs around the nation and the world, it is found that these MMORPGs are now moving to the community-oriented games, which emphasizes on the familiarity and communities of the game characters.

Despite of such enhancements in MMORPGs, the communication system between PC and NPC is still limited to the simple conversations. In this paper, the communication system of the existing MMORPG is named as 2D-Dialogue System, meaning that it is the two dimensional dialogue system based on the levels of PCs and the types of NPCs.

Fig. 2 shows the 2D-Dialogue system where the X axis of the graph is the levels of PCs and Y axis is the types of NPCs. In the games that employ the basic 2D-Dialogue systems, a PC always gets the same dialogues from a NPC no matter how many times the PC meets the NPC. There is no consideration of the relationship between the PC and NPC except that the NPC sometimes talks more politely depending on the level of the PC, which is far different from social communication in the real world.

3 3D-Dialogue System

In this paper, we attempt to extend the 2D-Dialogue system to the three dimensions by adding the familiarity between a PC and a NPC as shown in Fig. 3. The depth information of the 3D-Dialogue system corresponds to this familiarity. In the 3D-Dialogue system, the depth of the conversation describes the degree of additional and detail descriptions about the same event, which is represented as a Layer. 3D-Dialogue System allows game designers to implement several dialogue layers by the familiarity factor between a PC and a NPC into the 2D-Dialogue System.

The dialogue quantity and depth is determined by the PC's level and the type of NPC's personality, as well as the familiarity between the PC and NPC. This 3D-Dialogue system empowers the level-based communication. Since this system concerns the familiarity, which is an important element of the communication,

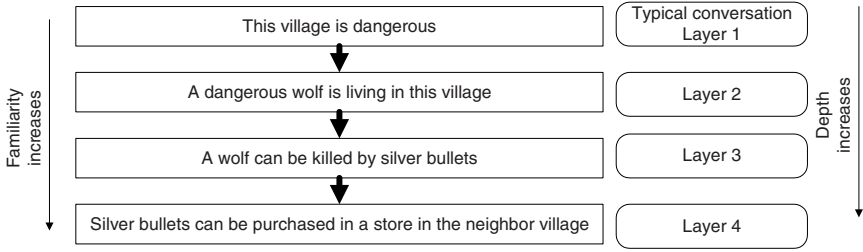


Fig. 4. Dialogues gets more detail as the familiarity increases

the new systems can increase the reality of the games and help the players feel more immersion because PCs are recognized and treated nicely by NPCs.

In the 3D-Dialogue System, X axis represents the grade of the i^{th} PC (PC_i), Y axis represents the types of the j^{th} NPC (NPC_j), and Z axis represents the familiarity between the PC_i and NPC_j . First, PC_i 's grade starts from the beginner's level defined by the game and gets promoted to a higher level as a user plays the game. Depending on the types, characteristics, themes, and contents of the game, the game designers should specify the types and extents of the PC's levels. Secondly, the kinds of NPC_j are defined by the game designers based on the roles and functions of the PWs. Finally, the familiarity shows how close PC_i and NPC_j are. In the real world, the acquaintance level between two people would be increased if the persons are encountered each other more often. The familiarity factor in the 3D-Dialogue system reflects this acquaintance level in the game environment. The familiarity factor can be affected by the following three main elements:

- (a) The number of meetings and conversations between the PC_i and the NPC_j
- (b) The number of events that are proposed by NPC_j and completed by PC_i
- (c) The number of times of PC_i giving out gifts or cash to the NPC_j

The value of familiarity in the 2D-Dialogue System games would be one no matter what the values of (a), (b), and (c) are. In other words, there is only one layer in these games. However, in 3D-Dialogue system games, the value of familiarity can be greater than or equal to two, indicating that the number of layers.

3.1 Information Exchanging Conversation in the 3D-Dialogue System

In 3D-Dialogue System, not only the base-level dialogue is provided to the user, but also additional talks are presented to the user throughout the layers based on the familiarity. That is, users get more descriptions and dialogues throughout the various layers so that they can understand the details of events or objects. Fig. 4 shows an example of the dialogues between a PC and a NPC that illustrates more complex conversations as the familiarity increases. In the 2D-Dialogue system

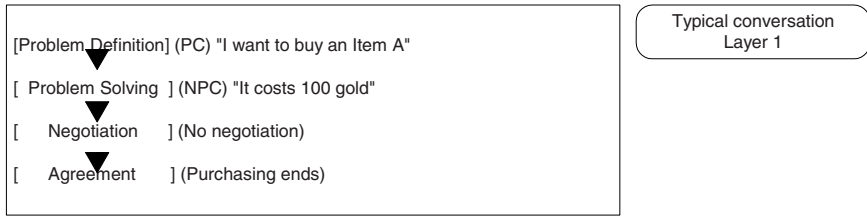


Fig. 5. Buying/Selling Conversation in a 2D-Dialogue System

game, it shows the talks from the first layer (Layer 1), whereas the layer of conversation is increased as the familiarity to the NPC gets higher.

As the dialogue system gets deeper, the layer provides more detail information about the topic of the conversation to the user. Fig. 4 illustrates more information that has been shown as the familiarity increases, in “the situation of a village” → “wolf” → “silver bullet” → “a store in the village next to this”.

3.2 Buying/Selling Conversation in the 3D-Dialogue System

According to a communication research, Dialogforschung, the conversation for buying and selling (hereafter just “buying/selling conversation”) is a kind of “komplexer dialog” [9]. It describes the steps of the “buying/selling conversation” deals as following: Problem specification → Problem solving → Negotiation → Mutual agreement. By using these four steps, we explore an example of buying/selling conversation in MMORPG.

As shown in Fig. 5, PC does not negotiate or make deals when it purchases items from a seller NPC. This is because it does not consider any familiarity between the PC and NPC. The conversation or purchasing process is not affected by the number of conversations they have had before, the number of completed events, or the gift offers.

Fig. 6 shows the conversation that employs the 3D-Dialogue System in the MMORPG. It shows the purchasing and dealing process of PC, which tries to negotiate the price of game items with a NPC in a store. This process shows that the chance of dropping price would be increased as the familiarity gets higher between the PC and NPC. In order to buy an item at cheaper price, PC needs to increase the familiarity to the NPC by making more conversations, giving a gift, or completing events.

4 Conclusion and Future Studies

MMORPG has been rapidly developed and improved with the wide availability of Internet service providers and high-speed network. Furthermore, game players desired more interesting online games with the advent of high quality hardware, 3D graphics, and software, and the game developers also have interests in applying innovative game designs. This paper explores the dialogue system, which is one of the elements of the game design.

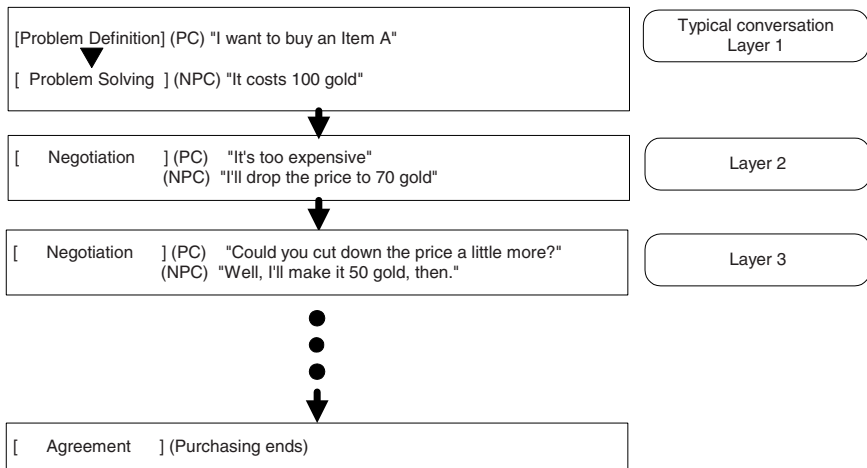


Fig. 6. Buying/Selling Conversation in a 3D-Dialogue System

This research proposes a 3D-Dialogue System that adds the familiarity in the 2D-Dialogue System. The 2D-Dialogue system only relied on the levels and types of PCs and NPCs, respectively. The 3D-Dialogue System allows game designers to develop a new era of online games. It would support more solid conversations, and evoke users' motivation and interests by allowing them to generate multiple layers of dialogues based on the familiarity factor between game characters. We believe this system would encourage more active participation of game players and the construction of the game flow because it is one of the important characteristics in games.

If this research is applied to the development of a MMORPG, then, it would create various interesting dialogue scripts for the game. In the future, if a MMORPG is developed with the research proposed in this paper, we will evaluate the game. It would be much easier to evaluate the system if we can build a game for a shrink-wrapped package or a console since the game often targets a single player, which would help speeding up the evaluation of the simulation and verification. The console game called "Toro and a day", published from Sony, was a great example that employed a measure of familiarity to the existing styles of adventure games.

The proposed 3D-Dialogue System allows diversifying the conversation. It also allows the construction of more realistic communities based on the diverse communication system. Thus, it would be applied to other multimedia digital contents based on the communication. For example, for some websites and mobile systems that rely on the textual contents could be benefited from this system by giving more confidence to the visitors of the environments. It would be also possible to automate customer support service or other reception work, which would decrease the dependency on the human resource.

In the future, we will also incorporate a psychology theory to study characteristics of various NPCs to make them more realistic. We will also explore the problems of designing various dialogue patterns to investigate the communication among game characters.

References

1. Mullidan, J., "Developing Online Games : An Insider's Guide", Pearson Education, pp 28,. 2003
2. <http://baram.nexon.co.kr/>
3. Korean Game Development & Promotion Institute, "2003 Korea Game White Paper", Korea Game Development & Promotion Institute, pp311,. 2003
4. <http://www.mabinogi.com>
5. <http://www.ragnarokonline.com>
6. <http://www.muonline.co.kr>
7. <http://www.lineage.co.kr>
8. <http://www.lineage2.co.kr>
9. Franke, Wilhelm., Entwicklung eines dialoggrammatischen Konzepts zur Beschreibung des Dialogtyps, In: Franz Hundsnurscher/Wilhelm Franke (Hg.): Das VERKAUFS-/EINKAUFGESPRACH. Eine linguistische Analyse. Stuttgart: Heinz Akademischer Verlag, pp. 76-108, 1985

A Hybrid Approach to Detect Adult Web Images

Qing-Fang Zheng, Ming-Ji Zhang, and Wei-Qiang Wang

Institute of Computing Technology, Chinese Academy of Science, Beijing, China
{qfzheng, mjzhang, wqwang}@jdl.ac.cn

Abstract. This paper presents a hybrid approach to discriminate benign images from adult images. Different from previously published works, our approach combines face detection and adaptive skin detection. First, face detection using haar-like features is performed, then skin color model is learned on-line by incorporating the information of color distribution in the face regions. Based on the result of face detection and adaptive skin detection, a set of semantically high-level features are extracted. These features have human-oriented meanings which can effectively discriminate adult images from benign images. Our two main contributions are i) a set of high-level features for adult or benign image classification and ii) a novel adaptive skin detection algorithm in still images. Experimental results are reported to demonstrate the strength of our approach.

1 Introduction

The Internet is one of the greatest inventions of all times, but it has also become a playground for pornographers. According to a commercial report [1], the number of pornographic web pages on the Internet in year 2003 has increased nearly 1800 percent compared with 14 million pages five years ago. Exposure to the sea of pornography can lead to many social problems including cyber-sex addiction. It is now an urgently necessary task to prevent people, especially children, from accessing this type of harmful material. A direct solution to Internet porn images filtering is to evaluate an image's content before displaying it. This paper presents a new approach to separate benign images from adult images by analyzing a set of semantically high-level features obtained from face detection and adaptive skin detection.

To our knowledge, there are no previously published works that used high-level features such as face to perform benign-or-pornographic images classification. Related works concerning this field mostly used low-level visual features. Wang et al. employed a combination of Daubechies' wavelets, normalized central moments and color histograms as a feature vector and matched it against a small number of features obtained from a training database [2]. Chan et al. used three simple features: the ratio of skin area to image area, the ratio of the largest skin segment to the image area and the number of segments in the image [3]. The features used by Jones and Rehg are [4]: percentage of pixels detected as skin;

average probability of the skin pixels size of the largest connected pixels; size of the largest connected component of skin; number of connected components of skin; percent of novel pixels, height and width of the image. In Image Guarder system develop by Zeng, a combination of skin feature, texture feature and shape feature are used [5]. Forsyth developed an algorithm that involved a skin filter and a human figure grouper to find naked people in the image [6].

Although low-level features are easy to compute, they are insufficiently accurate because of semantic gap: Human's interpretation of image's content as pornographic or not is so abstract that there is no simple computational transformation that will map low-level image features to human perception. We propose to use high-level features to achieve pornographic or non-pornographic image classification. The features we used are all related to human face. Our approach takes advantage of research achievements of face detection technique because face detection is now being extensively studied and some effective and efficient face detectors have been introduced in the literature. There are two main contributions of our work. The first is a set of semantically high-level features for image classification and the second is a novel adaptive skin detection algorithm in still images which is robust to illumination conditions and not biased by human races.

The rest of the paper is organized as follows. In section 2, we will detail our hybrid approach including face detection, on-line skin color modeling, adaptive skin detection, feature extraction and image classification. In section 3, we will provide experimental result. Conclusion and future directions will be discussed in section 4.

2 Our Hybrid Approach

In this section, we describe our approach to separate benign images from pornographic images by analyzing image content. We will start with the overall architecture, followed by a detailed discussion of the main components.

2.1 General Overview

The basic process flow of the proposed approach is as follows. First, face detection is performed on the input image using a set of over-complete haar-like features. Once face is detected, color distribution of face region is computed and used for on-line human skin color modeling. And then adaptive skin detection is performed on the whole image. Based on the result of face detection and skin detection, a set of semantically high-level features are extracted and fed into a decision tree classifier. The whole procedure can be visualized in Fig. 1.

2.2 Semantically High-Level Features

High-level features gain advantages over low-level features in that they narrow down the semantic gap between human perception and raw image data. To interpret an image's content, we must identify some important features or objects

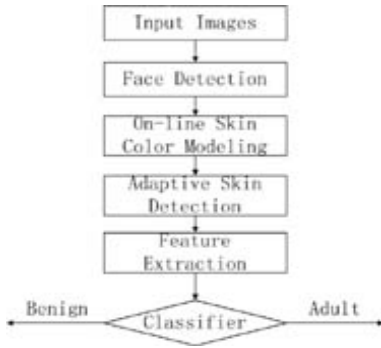


Fig. 1. Overall process of our approach



Fig. 2. Face detection, the red circle marks the detected face region.

in the image. Face is the most distinctive part of human body. Detection of faces can allow the observer to form a hypothesis about the presence of humans in the scene and other measures can be taken to verify whether or not they are nudes. We empirically choose five features for benign-or-pornographic image classification. These features are all related to face:

(1) Face Number: the number of faces in the image. Pornographic images usually do not contain too many faces. We define $F1$ as:

$$F1 = \text{number of face in the image} \tag{1}$$

(2) Face Area: the area of face regions. Images in which face regions cover too much area may be full-face portraits. We define $F2$ as:

$$F2 = \frac{\text{number of pixels in face region}}{\text{number of pixels in whole image}} \tag{2}$$

(3) Face Position: Images which present face in the center are usually benign. We define $F3$ as distance between the center of face region and the center of image normalized by the minimum length of image axis:

$$F3 = \frac{\text{distance}(\text{center}_{\text{face}}, \text{center}_{\text{image}})}{\min(\text{image.width}, \text{image.height})} \tag{3}$$

(4) Skin Ratio I: the ratio of whole image skin area to face area;

(5) Skin Ratio II: the ratio of the largest skin segment to the face area. The higher the Skin Ratio I and Skin Ratio II are, the more likely the images are pornographic. We define $F4$ and $F5$ as follows:

$$F4 = \frac{\text{number of skin pixels in whole image}}{\text{number of pixels in face regions}} \tag{4}$$

$$F5 = \frac{\text{number of skin pixels in largest skin segment}}{\text{number of pixels in face regions}} \tag{5}$$

2.3 Face Detection

Face detection plays an important role in the our image classification approach. Our face detector is based largely on the work of Paul Viola [7]. A cascade of boosting classifiers is built on an over-complete set of haar-like features. In each stage of the cascade, a variant of AdaBoost is used to integrate the feature selection and classifier design in one boosting procedure. Only positive result from the previous classifier needs further more complex evaluation. By adopting this simple-to-complex strategy, most non-face candidates are rejected in earlier stage of cascade with little computation costs. Details of the face detection algorithm can be found in [7]. In order to successfully detect rotated faces, each input image is passed to face detector three times. The first time is the original image, the next two times are its rotated variants with rotation angle of 45 degrees anticlockwise and clockwise respectively. We do this because we find that faces in adult images are usually within a rotation angle of no more than 90 degrees from vertical direction. Fig. 2 illustrates our face detection method.

2.4 Online Skin Color Modeling

Skin detection is an important technique for identifying adult images because of the fact that there is a strong correlation between image with large patches of skin and adult images. However, accurate skin detection is a non-trivial task. In traditional skin detection schemes, very often a static skin color model is learned off-line and each image pixel is checked whether or not its color value satisfies the learned model. Skin color varies greatly between different human races. To make things worse, skin color, as measured by camera, can change when illumination condition changes. Therefore, skin detection that uses a static skin color model is certain to fail in unconstrained imaging conditions. Here we propose a novel adaptive skin detection method based on the result of face detection. Once face is detected, color distribution in face region is used as useful context information for on-line skin color model building. The proposed method is robust to imaging conditions and not biased by human ethnicity.

Our skin detection method takes advantage of the fact that the face and body of a person always share same colors. Color distribution of face regions can provide useful cues to detect skin regions of other body parts. We choose to build skin color model on YCbCr color space and use a normal distribution $N(\mu, \sigma)$ to represent the distribution of each skin color component (Y,Cb,Cr). Color values of image pixels in face region are viewed as an ensemble of skin color samples:

$$\Omega = \{\{y_1, cb_1, cr_1\}, \{y_2, cb_2, cr_2\}, \dots, \{y_K, cb_K, cr_K\}\} \quad (6)$$

Then the mean and variance of each normal distribution can be computed. For distribution of Y component:

$$\mu_y = \frac{1}{K} \sum y_j \quad \sigma_y^2 = \frac{1}{K-1} \sum (y_j - \mu_y)^2 \quad (7)$$

K is the size of Ω , i.e., number of pixels in face regions.

The distribution parameters of Cb and Cr components can be computed similarly. Pixel outside of face region is classified as skin pixel if it satisfies the following requirements:

$$\|y - \mu_y\| \leq a_y \sigma_y \quad \text{and} \quad \|cb - \mu_{cb}\| \leq a_{cb} \sigma_{cb} \quad \text{and} \quad \|cr - \mu_{cr}\| \leq a_{cr} \sigma_{cr} \quad (8)$$

y, cb, cr are the color values of the pixel to be classified and a_y, a_{cb}, a_{cr} are threshold values to be adaptively determined which will be described in section 2.5.

2.5 Adaptive Skin Detection

We do not use fixed threshold values because through experiment we find fixed thresholds can not separate skin region from non-skin region which have colors similar to skin. Instead, we select threshold values adaptively by taking the texture property of human skin region into consideration. Skin region is usually homogeneous and has smooth texture. Using texture characteristics of skin region to find an optimal threshold for skin segmentation was proposed by Phung [8]. Two main aspects can differentiate Phung's method from ours. One is the fact that Phung's method performs on skin score map computed from Bayesian decision theory while our method performs on original images, the other is the region homogeneity measures. The homogeneity measures we used are:

$$\sigma_{region}^y < 0.5 \mu_{region}^y \quad \text{and} \quad \sigma_{region}^{cr} < 0.4 \mu_{region}^{cr} \quad \text{and} \quad \sigma_{region}^{cb} < 0.4 \mu_{region}^{cb} \quad (9)$$

σ_{region}^y is the standard deviation of Y color component in the region, μ_{region}^y is the mean of Y color component in the region. $\sigma_{region}^{cr}, \mu_{region}^{cr}, \sigma_{region}^{cb}, \mu_{region}^{cb}$ have similar meanings.

Firstly, each image pixel is classified using equation (8) with initial threshold values $a_y = 2.5, a_{cb} = a_{cr} = 2.0$. After this coarse skin detection, we can get some skin regions. For each large enough region, we check the homogeneity property. If it is homogeneous, it is considered as true skin region. If it's not homogeneous, the threshold values are decreased by a factor of 0.9 respectively, and skin detection using the new threshold values is performed on the region. This process continues until the skin regions become homogeneous. Fig. 3 gives some experimental result. The top row is coarse skin detection result, the bottom row is the final result. The original images are not presented for offensive reasons.

When more than one faces are detected in the image, each face region is used to construct a skin color model, and each model is used to perform skin detection on the whole image. The final result is a logical OR combination of each of the detected regions obtained respectively with each skin color model.

2.6 Image Classification

After face detection and skin detection, high-level features described in section 2.2 can be easily extracted. We used these features to do pornographic-or-not images classification. Although there are many sophisticated classification methods



Fig. 3. Skin detection result. The top row are coarse skin detection results, the bottom row are the final refined results. For offensive reasons, original images are not presented here.

such as Support Vector Machine (SVM) and Artificial Neural Network (ANN), we opt for using decision tree[9] for its simplicity and its particular efficiency when our features have human-oriented meanings. Fig. 4 shows the tree we used. Nodes in the tree involve testing a particular feature by comparing its value with a constant threshold. The tree calls firstly for a test on $F1$, and the first two branches correspond to the two possible outcomes. If $F1$ is bigger than a threshold, the outcome is benign. If outcome is the other branch, a second test is made, this time on $F2$. Eventually, whatever the outcome of the tests, a leaf of the tree is reach that dictates the classification result.

3 Experimental Results

This section reports testing result of our approach. The performance of our approach is tested on a database of 2196 images, among which 451 images are manually labeled as adult images. Benign images are 1119 human images and other 626 images including animal, plant, landscape images. Adult images and human images are downloaded from the Internet. Other images are from Corel image library. It should be pointed out that adult images and human images we used all contain frontal or near-frontal view faces because our face detector currently can not detect faces with too much out-of-image-plane rotation. The test result is shown in Table 1. For offensive reasons, we do not list the adult images here. Fig. 5 gives some images that successfully detected as benign images

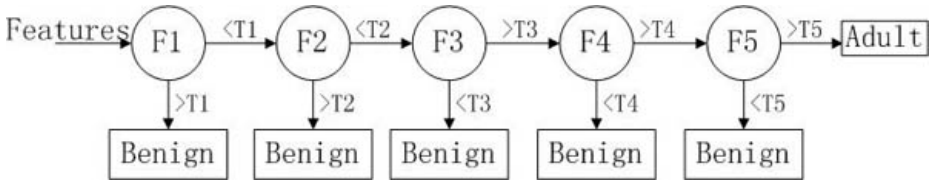


Fig. 4. Schematic depiction of image classification process. T_1, T_2, T_3, T_4 and T_5 are predefined thresholds.

Table 1. Testing Result

Images	Detected as Adult Images	Detected as Benign Images
Adult Images (451)	412 (91.35%)	39 (8.65%)
Benign Human Images (1119)	99 (8.47%)	1020 (91.53%)
Other Benign Images (626)	36 (5.75%)	590 (94.25%)

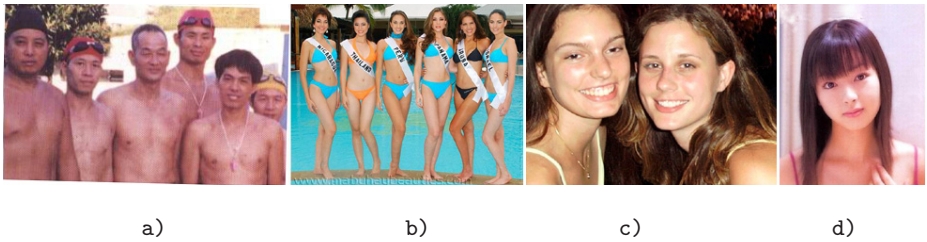


Fig. 5. Some example images successfully detected as benign images because many faces are detected in a) and b), c) contains large area of face region and d) present face in the center. These images can be wrongly classified as adult images only using low-level visual features.

by our method. These images can be easily classified as pornographic images using low-level features, because these images present large area skin patches.

4 Conclusion and Future Directions

In this paper, we propose an approach to automatically discriminate benign images from adult images. Our approach takes advantage of achievements of face detection research. Based on the result of face detection and adaptive skin detection, our approach achieves image classification by analyzing a set of semantically high-level features. Face detection plays a dominant role in our approach. In future work, we will focus on more robust face detector. And we will pay attention to the combination of low-level and high-level visual feature to more effectively detect adult images.

Acknowledgement. This work has been financed by the National Hi-Tech R&D Program (the 863 Pro-gram) under contract No.2003AA142140. The authors want to thank Wei Zeng for his constructive suggestions.

References

1. N2H2 Inc. N2H2 Reports Number of Pornographic Web Pages Now Tops 260 Million and Growing at an Unprecedented Rate. [Http://www.n2h2.com](http://www.n2h2.com).
2. Wang, J., Wiederhold, G., Firshein, O.: System for screening objectionable images using Daubechies' wavelets and color histograms. IDMS'97. Volume 1309, Springer-Verlag LNCS (1997), pp. 20-30
3. Chan, Y., Harvey, R., Smith, D.: Building systems to block pornography. In Eakins, J., Harper, D., eds.: Challenge of Image Retrieval, BCS Electronic Workshops in Computing series (1999), pp.34-40
4. Jones, M.J., Rehg, J.M.: Statistical Color Models with Application to Skin Detection, Technical Report, Cambridge Res. Lab., Compaq Comput. Corp., 1998
5. Zeng, W., Gao, W., Zhang, T., Liu, Y.: Image Guarder: An Intelligent Detector for Adult Images. Asian Conference on Computer Vision. ACCV2004, Jeju Island, Korea, Jan.27-30,2004, pp198-203
6. Fleck, M.M., Forsyth, D.A., Bregler, C., Finding naked people. ECCV. Volume II Spring-Verlag, (1996), pp.593-602
7. Viola, P., Jones, M. J.: Rapid Object Detection Using A Boosted Cascade of Simple Features. IEEE Conference on Computer Vision and Pattern Recognition, Jauai, Hawaii, 2001
8. Phung, S.L., Chai, D., Bouzerdoun, A.: Adaptive Skin Segmentation in Color Images. ICASSP 2003
9. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools with Java Implementations, Morgan Kaufmann, San Francisco, 2000

A Hierarchical Dynamic Bayesian Network Approach to Visual Tracking

Hua Li^{2*}, Rong Xiao¹, Hong-Jiang Zhang¹, and Li-Zhong Peng²

¹ Microsoft Research Asia, No.49 Zhichun Road, Haidian District,
Beijing 100080, P. R. China
{t-rxiao,hjzhang}@microsoft.com

² LMAM, Department of Mathematics, School of Mathematical Sciences,
Peking University, Beijing 100871, P. R. China
lihua@math.pku.edu.cn, lzpeng@pku.edu.cn

Abstract. Usually a uniform observation strategy will result in frustrated tracking processes. To address this problem, we construct a flexible model with Hierarchical Dynamic Bayesian Network by introducing hidden variables to infer the intrinsic properties of the state and observation spaces. With this model, a dynamic-mapping is built between target state space and the observation space. Based on a decoupling based inference strategy, a tractable solution for this algorithm is proposed. Experiments of human face tracking under various poses and occlusions show promising results.

Keywords: Visual tracking, dynamic bayesian network, sequence monte carlo.

1 Introduction

Visual tracking plays a crucial role in many multimedia applications such as video surveillance (individual, vehicle tracking), communications (low bit-rate video coding and object-oriented MPEG-4) and intelligent human computer interfaces. In these applications, face position from tracking is preliminary requirement for successive procedure.

In general, visual tracking technique involves four basic factors: target representation, state prediction, observation obtaining (feature measurement) and state estimation. Each of these factors is paramount to achieve a successful tracking process and makes a sub-problem of visual tracking.

Target representation, concerning target's geometry property, motion state, texture information, and etc., is critical to discriminate the target from both the similar objects and the clutter background. It forms a fundamental problem in computer vision. Many works have been published during last decades. For instance, active shape models [4][6], and skin color distributions [1][9][12] are mostly applied to describe target appearance. In [5][11], a more robust description, both shape structure and color information are employed. However, these

* This work was performed at Microsoft Research Asia.

appearance based approaches are suffered from target occlusions (including self-occlusion caused by pose variation). Moreover, even partially occlusion occurs, the appearance representation will be changed dramatically from the normal template. And this results in unstable tracking based on motion difference [8].

Both state prediction and estimation rely on certain statistical models, such as Particle Filter [3] and Maximum Likelihood learning [2]. A labeled training database is also necessary for learning the parameters and the models' structures are usually assumed in advance for simplicity. As we can see, the observation model is closely related to target representation. For example, target's contour edge is mostly expected to be detected when the target is represented by shape model. To cope with non-stationary clutter and other object's occlusions, [10] proposed a switching observation based models. Still, the tracker will be lost when a similar object appears in the clutter or the target is occluded by itself. Thus, to construct an appropriate combination of target representation and observation strategy is crucial to form a both ideal and effective statistical model. That is, the observation strategy should adequately discover the intrinsic property of target's current state, which helps discriminate the current state from other objects' and also other states of itself.

From the point of view of Mathematics, a visual tracking problem comes down to establishing an accurate dynamic-mapping between the target's state space and the observation space. Both spaces are continuous ones with high dimensionality, which are hard to be modeled directly. In previous works, a first-order or second-order Hidden Markov process is usually applied to model the dynamic between immediate time steps [4], which in fact is a "a linear construction + a Gaussian process diffusion" frame. The important based sampling strategy aims to approximating the observation space using discrete sequence data [3]. Usually, in existent algorithms these two processes run separately. However, as a matter of fact, the couple of spaces are not independent from each other but closely interrelated. To make the most of this coupling information should be a better strategy.

In this paper, we construct a flexible model with Hierarchical Dynamic Bayesian Network by introducing hidden variables to infer the intrinsic properties of the state and observation spaces. With this model, a dynamic-mapping is built between target state space and the observation space. Based on a decoupling based inference strategy, a tractable solution for this algorithm is proposed.

This paper is organized as follows. In section 2, we present a hidden variable introduced state space model. A dynamic observation model will be described in section 3. Then, we give a mathematical presentation on the inference of the proposed Hierarchical Dynamic Bayesian Network in section 4. Section 5 contains experimental results and analysis. Section 6 concluded the paper and suggested future work.

2 Hidden Variable Introduced Dynamic State Space Model

Denoting the target state at time t by X_{t-1} we have the following first-order Markov chain based state transition model:

$$x_t = D_{dynamic}(t) * X_{t-1} + W_{eight} * G_{aussian}(t) \tag{1}$$

It is assumed that the new state is conditioned directly only on the immediately preceding state. $G_{aussian}(t)$ is the process noise with a Gaussian distribution $N(0, Q(t))$. When the target performs typical motions, $D_{ynamic}(t)$ and $G_{aussian}(t)$ are usually estimated from training data using Maximum Likelihood Estimation [2]. However, when the state is highly time-varying, there will be problems even with a higher order model. Here we introduce a hidden state variable S_t :

$$X_t = D_{ynamic}(S_t) * X_{t-1} + W_{eight} * G_{aussian}(S_t) \tag{2}$$

For reasonable simplification, we set S_t belong to a discrete space and assume that the dynamic of S_t follows a first-order Markov process. With sufficient training data, $\Theta = s_0, s_1, \dots, s_m - 1$ can be selected to span the state space statistically, using the learned conditional probability distribution $P(s_i | s_j)$ and the prior distribution $P(s_i)$. Then, the distribution of S_t can be represented as:

$$P(S_t) = \left(\prod_{k=1}^{t-1} P(S_k = s_i | S_{k-1} = s_j) \right) * P(S_0) \tag{3}$$

Denote $T_{transition}(i, j) = P(S_t = s_i | S_{t-1} = s_j)$, $i, j \in 0, 1, \dots, M - 1$. When $i = j = 0$, it gives an initial state distribution.

Especially, it should be noticed that S_t is just an abstract symbol used to analysis the complex state space. Its style is not unique to a specific state space and the character of its value space also depends on the different property of the target's dynamics.

We depict the state space using the following dependency graph:

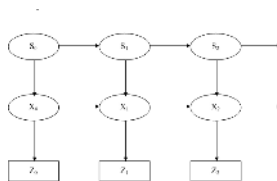


Fig. 1. The hidden state variable introduced dynamic Bayesian network. Arcs show dependencies between variables.

Directly inferred from the graph, we have the following propagation:

$$P(X_t, S_t | \bar{Z}_t) \propto P(Z_t | S_t, X_t) * \int P(S_t, X_t | S_{t-1}, X_{t-1}) * P(d(S_{t-1}, X_{t-1}) | \bar{Z}_t) \tag{4}$$

3 Flexible Dynamic Observation Model

In common, a time-invariant observation model or a switching observation model [10] is used to approximate the conditional probability distribution $P(Z_t|S_t, X_t)$. Comparing with the state space model, a fixed observation is not a best strategy to reflect the different property of various target state. For example, when we use a set of measurement points to collect face feature under various poses:



Fig. 2. Labeled feature points shown in a 3D face model.

The labeled face feature points show different appearance comparing with not only itself in different poses but also others in the same pose. In details, to the tracker, the feature points around eyes give more robust information about pose variable, but less sensitive to dynamic. The ones on nose show great sensitivity to head pose variance. Even a feature point which is invisible in some pose provides certain useful information. It will be a more flexible and accurate solution to get a credible likelihood expression if we give different weight to different feature points under various face poses. Namely, we use this strategy to give an analytic representation of the observation space through statistical learning.

A hidden observation variable U_t is introduced to describe the utility of the feature points. It imposes additional structure in the observation space. The following hierarchical dynamic Bayesian network depicts the whole tracking process:

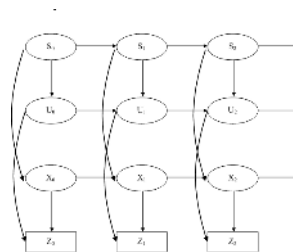


Fig. 3. Hierarchical dynamic Bayesian network with extra structures imposed in both state space and observation space.

Given the dynamic Bayesian network in Fig.3, the aim of tracking is to inference the distribution $P(X_t, S_t, U_t | \bar{Z}_t)$. Denoting $\omega_t = (X_t, S_t, U_t)$, we have:

$$P(\omega_t | \bar{Z}_t) \propto P(Z_t | \omega_t) * \int P(\omega_t | \omega_{t-1}) * P(d(\omega_{t-1} | \bar{Z}_t)) \tag{5}$$

It is advisable to learning the conditional probability distribution $P(U_t | U_{t-1})$ from training data using approximation techniques.

4 Inference in Hierarchical Dynamic Bayesian Network

The goal of inference in complex dynamic Bayesian network is to find a tractable posterior probability distribution of the hidden states of the system given some known sequence of observations \bar{Z}_t .

Sequence Monte Carlo methods [3] also known as Particle filters, is popular in recent years as a numerical approximation for complex models. The basic idea of Particle Filter is to get updated samples generated from an appreciately chosen proposal distribution which depends on the different utility of old ones. As the distribution of U_t with flexible dynamic is hard to model, we designed a decoupling based structured inference algorithm to approach a tractable calculation.

4.1 Decoupling Based Structured Inference

Given the observed data sequence \bar{Z}_t and the model parameters $\omega_t = (X_t, S_t, U_t)$, the objective is often defined as minimizing the Kullback-Leibler (KL) divergence between the approximate and the true posterior distribution:

$$\begin{aligned} \log(P(\bar{Z}_t)) - LB(\log(P(\bar{Z}_t))) &= - \int R(d\omega_t) * \log(P(\omega_t | \bar{Z}_t) \setminus R(\omega_t)) \\ &= KL(R(\omega_t) || P(\omega_t | \bar{Z}_t)) \end{aligned} \tag{6}$$

R denotes the approximation of the posterior distribution. Then for the computationally efficient inference, we define R by decoupling the baroque dynamic Bayesian network shown in Fig.4.

If we take U_t as a parameter, its optimal value can be obtained by minimizing the KL-divergence w.r.t U_t . Since the two sub-graphs are inter-independent from each other, we define R as follows:

$$R(X_t, S_t, U_t | \bar{Z}_t) = R_x(X_t | U_t, Z_t) * R_s(S_t | U_t). \tag{7}$$

This leads to a Hidden Markov process $R_s(S_t | U_t)$ and a linear dynamic system $R_x(X_t | U_t, Z_t)$. A LDS inference is used to obtain $\langle X_t \rangle = E[\bar{Z}_t]$, also, an inference in input-output HMM is applied to estimate $\langle S_t \rangle = P(S_t | \bar{U}_t)$, where \bar{U}_t denotes the sequence of U_t .

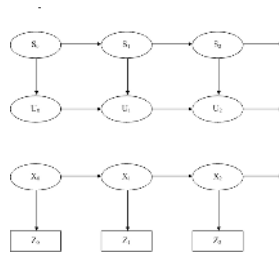


Fig. 4. The decoupled graph of the densely connected Bayesian network. It reduces the original one into two separate graphs.

The whole algorithm can be summarized as follows:

Generate $X_t^n, U_t^n, S_t^n, \Pi_t^n$ from Z_{t-1}^n, Π_{t-1}^n :

- (1) update samples based on $\{\Pi_{t-1}^n\}$, get $\{X_t^n, S_t^n, \Pi_t^n\}$. (LDS inference)
- (2) estimate $\{U_t^n\}$ from $\{X_t^n, S_t^n, \Pi_t^n\}$.
- (3) estimate $\{S_t^n\}$ from $\{U_t^n\}$ (HMM inference).
- (4) correction : calculate $\Pi_1^n = P(Z_t|X_t^n, U_t^n, S_t^n)$

4.2 Learning Maximum Likelihood

An extended EM technique is applied to learn the DBN parameters. The decoupling based structured inference constructs the Expectation process. Parameter update equations in Maximization process are obtained by maximizing $E(P \log(X_t, S_t, U_t, Z_t))$ using the sufficient statistics from inference process.

5 Experimental Results

5.1 Tracking Under Various Poses and Occlusion

We use the 3D head model show in Fig.2 to get training data reflecting the utility of feature points in multi-view.

The result for the real single face tracking with multi- view is shown in Fig.5. With the changing of head pose, the “most” (approximately) appropriate U_t was calculated to give corresponding weighs to each feature group. 900 particles are used to obtain the tracking result.

With the occlusion, almost the same particles are needed to get the accurate tracking result. When a feature group’s confidence is going down to a low level, the sampling number of following step in this part is reduced. At the some



Fig. 5. Single face is tracked under various poses.



Fig. 6. Single face with occlusion in multi-view is tracked.

time, the rising sampling number of high-confidence-level feature group keeps the robust of the proposed algorithm. Fig.6 shows the corresponding experimental result.

6 Conclusions and Future Work

Different target state contains different property which is usually smoothed by a fixed observation model. Though, a switching observation model is more flexible and robust, as we all know, it is impossible to model all kinds of clutters and occlusions into a switching observation model. Also, it is not the case that the more feature points/sampling number is definitely the better, since too much information may cause over-fitting and will result in deviated decisions of a tracker. The advisable strategy is to model the target itself of various states. This is why we introduced the utility variable of observation and also the corresponding hidden state variable.

The future work will focus on tracking faces with various illumination and expressions. Analyzing the intrinsic corresponding between the two hidden variables is probably a promising direction.

References

- [1] Dorin. Comaniciu, Visvanathan. Ramesh, and Peter Meer: Real-time Tracking of Non-rigid Objects using Mean Shift, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, 2000, volume II, pp 142-149.
- [2] A. P. Dempster, N. M. Laird, and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Society Series B*, 39:1-38, 1997.
- [3] A.Doucet, J.Godsill, and C.Andrieu: On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing*, volume 10, no.3 ,pp 197-208,2000.
- [4] Michael. Isard, and Andrew. Blake: Contour Tracking by Stochastic Propagation of Conditional Density, In Proc. of European Conf. on Computer Vision, Cambridge, UK, 1996, pp 343-356.
- [5] Michael. Isard, and Andrew. Blake: ICONDENSATION: Unifying Low-level and High-level Tracking in a Stochastic Framework, In Proc. of European Conf. on Computer Vision, Freiburg, Germany, 1998, volume I, pp 767-781.
- [6] Michael. Isard, and Andrew. Blake: Condensation-Conditional Density Propagation for Visual Tracking, *Int.J Computer Vision*, 29,1,5-28, 1998.
- [7] V. Krishnamurthy and J. Evans: Finite-dimensional filters for passive tracking of markov jump linear systems. *Automatica*, 34(6):765-770, 1998.
- [8] Valdimir. Palvovic, James M. Rehg, and Tat-Jen Cham: A Dynamic Bayesian Network Approach to Figure Tracking using Learned Switching Dynamic Models, In Proc. IEEE Int'l Conf. on Computer Vision, Corfu, Greece, Sept. 1999, volume I, pp 94-101.
- [9] Y. Raja, S. McKenna, and S. Gong: Colour Model Selection and Adaptation in Dynamic Scenes, In Proc. of European Conf. on Computer Vision, Freiburg, Germany, 1998, volume I, pp. 460-475.
- [10] Ying. Wu, Hua. Gang and Ting. Wu: Switching Observation Models for Contour Tracking in Clutter, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Madison, WI, July, volume I, pp 295-302.
- [11] Ying. Wu, and Thomas. S. Huang: A Co-inference Approach to Robust Visual Tracking, In Proc. of IEEE Int'l Conf. on Computer Vision, Vancouver, Canada, 2001, volume II, pp. 26-33.
- [12] Ying. Wu, and Thomas. S. Huang: Color Tracking by Transductive Learning, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, 2000, volume II, pp 133-138

A Hybrid Architectural Framework for Digital Home Multimedia Multi-modal Collaboration Services

Doo-Hyun Kim¹, Vinod Cherian Joseph², Kyunghee Lee², and Eun Hwan Jo¹

¹ Konkuk University, Seoul, Korea
{doohyun, ehjo}@konkuk.ac.kr
<http://kkucc.konkuk.ac.kr/~doohyun>
² ETRI, Taejon, Korea
{vinod, kyunghee}@etri.re.kr
<http://www.etri.re.kr>

Abstract. Home Server plays an important role to manage all home network devices and multimedia content services in the digital home. Multimedia collaboration over home networking devices is inhibited by several factors. The decision based on a single input agent to serve all the needs of the home users is to be enhanced to facilitate multi-modal inputs with multiple integrated User Agents working together to facilitate the new generation of dynamic home networking devices. Display has to be adaptive and context sensitive to adapt to the dynamic needs of the user. In this paper, we present a hybrid architectural framework as our recent efforts for extending the functionalities of Home Server in order to support multi-modal context-sensitive multimedia collaboration services.

1 Introduction

Recently, according to the rapid enhancement of the digital home technologies and services, Home Server[1,2] is expected to play the pivotal role of managing all home network devices and multimedia content services through home networks in the digital home. Home Server can be used as television with any digital monitor, surveillance equipment with a camera attached to the Home Server, messaging server as well as a streaming server, a data mining server or a home controller based on the needs of the home user. The Home Server was developed on the ETRI flavor of the embedded linux kernel, Qplus, that is optimized for multimedia processing.

The array of devices envisioned to operate in the digital home include mobile handsets, Personal Digital Assistant (PDA), other wireless devices on the wireless local area network (WLAN), bluetooth network and other wireless personal area network (WPAN), wireless sensor networks and home automation devices in addition to existing home server accessories. The plethora of possible inputs from diverse devices such as a camera, microphone, sketch recognizer, gesture recognizer, sensors and/or text have to be computed in a hybrid fashion to obtain the best possible output action for a particular sequence of inputs relative

to time and other parameters. Also, output has to dynamically change based on changes in user context. So display has to be adaptive and context aware to the needs of the user[3,4].

Managing multi-modal inputs in a dynamic context sensitive home environment from a plurality of inputs necessitates changes to existing design of home servers. Mobile devices need to be enhanced to support mobile agents and collect a multitude of inputs for user input, localization and security context recognition. This paper proposes the design methodology to accept multi-modal user inputs in the digital home and proposes a hybrid collaborative architecture to process the dynamic adaptive inputs with learning system design for the best output action based on context. It provides a probabilistic inference mechanism and dynamically adapts the output to the needs of the user input and security context.

1.1 Related Works

Hybrid multi-modal analysis is a hot topic for the networking world and several distinguished researchers are constantly striving to obtain a breakthrough to correctly model the dynamic network world. Suresh Jagannathan et al.[3] have modeled several distributed object models for hybrid architectures as demonstrated in “Transparent Communication for Distributed Objects in Java”. Their work primarily deals with the design of a software agent code and its location in a networked communication system. Their work identifies an object task, agent ID and task stack and defines the migration of the agent on a run-time system to the host. They define the software architecture of the object memory and classes for distributed agents operating network-centric applications. Our applications are not network-centric and are relative to the operation of an ad-hoc/fixed mobile digital home environment. Our design involves the methods for dynamic context-sensitive applications to adapt to the varying home environment conditions that accept multi-modal mobile inputs. The output is determined in a probabilistic context-sensitive architecture that adapts dynamically to changing security contexts, user profile settings and other predefined user parameters.

Abbot et al.[4] have defined a method, system, and computer-readable medium described for dynamically determining an appropriate user interface (UI) to be provided to a user in their work “Model-based synthesis of complex embedded systems” and related work on dynamic user interfaces. Their work facilitates the user of a wearable computing device to dynamically display one of the several stored user interfaces appropriate to the context. It defines their method to display a dynamic user interface based on predefined contexts and uses a software engine to select the appropriate UI. Our work involves accepting simultaneous multi-modal inputs from several devices configured to operate in the home environment and processing the different inputs to obtain the best output and configure the output to be displayed based on the user context. Our hybrid collaboration engine uses probabilistic inference based on individual users’ choices and previous usage patterns.

This section illustrates the introduction and work relative to the paper. The subsequent section presents the role of the multimodality in multimedia col-

laborations. Section three provides a hybrid architectural framework for multi-modality in collaboration. The conclusion section concludes the paper with pointers to further research.

2 Multi-modality in Multimedia Collaboration

2.1 Home Server and Multimedia Collaboration

The basic wireless multimedia collaboration architecture of Home Server is as shown in Fig. 1[2]. The Home Server has the V2oIP(voice and video over IP) video-conferencing stack that serves as proxy server for all calls to home networking devices and also acts as a gateway for devices with limited capability to transfer video. Some devices may need transformation from QCIF to Sub-QCIF format and image, video and protocol transformation occurs at the Home Server. The Home Server applications also include T.120 compliant whiteboard, screen sharing, and co-browsing for tele-education, web-based call center, and/or remote healthcare services[5,6]. The current design of Home Server facilitates a plethora of single input home conferencing applications. The users connect to the remote user over telephone or voice activation to dial to the remote IP address using V2oIP over SIP(Session Initiation Protocol)[7].

Wireless LAN is the only perceived home network that supports true multimedia data over wireless and other home networks currently support multiple media data. Streaming multimedia is effectively presented to the user by buffering and network bandwidth optimization. Multimedia data is sent over RTP/UDP[8] using robust header compression techniques and 40% of the WLAN bandwidth is reserved for multimedia. This provides 2.67Mbps per channel and facilitates a practical capacity of 3 simultaneous multimedia channels for the digital home at any time instant. The maximum available capacity per user is 2.67Mbps on a single channel and this can exceed to 3*2.67Mbps based on network load and channel usage criteria.

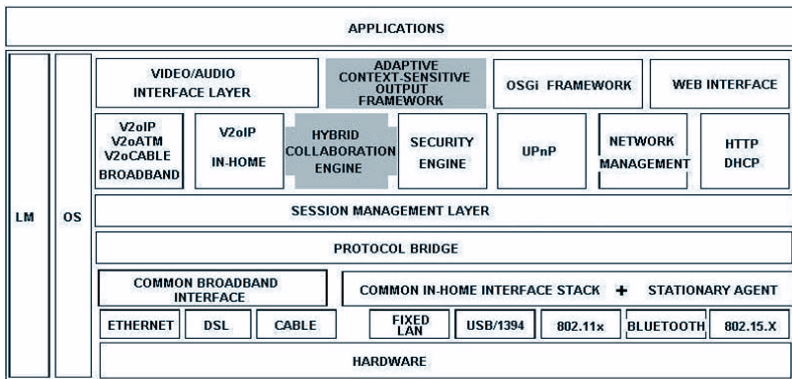


Fig. 1. Home Server Architecture for Multimodal Processing

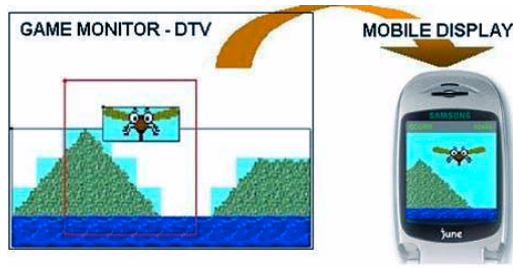


Fig. 2. Multimodality in Gaming Application

2.2 Multi-modality Issues

Multi-modality is critical to multimedia collaboration in wireless devices. Multi-modality facilitates bandwidth savings and usage of the limited wireless display to focus on the critical changes in multimedia data facilitating home users with clearer picture views for gaming, monitoring, remote surveillance and other applications. The usage of multi-modal inputs also facilitates clearer analysis of wireless multimedia data and inputs to obtain the best possible output for a series of inputs.

Suppose that the user wants to perform a voice query on google for “windows”. This term is generic and does not facilitate the system to provide the best output envisioned by the user. The addition of context related data and other multi-modal inputs from camera, gesture and other input devices facilitate the system to learn that the user intends to find the output for the wooden framework of the home window and his prior activities stored in the input block set provide more accurate analysis of what the user wants precisely. The Single-Input Single-Output(SISO) system would directly lead you to several search results for Microsoft Web URL which is frustrating for a Smart-Phone user with limited display to scroll down to the relevant search with excellent vision and patience. Multi-modality in this context allows the user to continuously provide sequential inputs to streamline the search to his relevant output. Let us consider the case scenario of a gaming user on a stringent display device like mobile phone or other proprietary gaming device. The usage of our architecture on the mobile device allows the user who tries to shoot the eye of the bird at the exact instant to correctly emulate the timing. This sort of experience as illustrated in Fig. 2 makes gaming experience truly fascinating to the game user and helps children to have a better vision with lesser strain on eyes with the addiction trends of our younger siblings to wireless devices.

Let us consider the case of security context for multi-modality in action in a digital home. Child lock on television sets is an important feature of our home to protect the disposal of explicitly sexual material to the younger siblings. When the elderly person watches a movie on a projector connected to the home server, he cannot control the display of content on the single input systems. Our sys-

tem stores user profiles and the user manager agent in collaboration with the security agent displays context-based output. Our system adapts to environment changes in home and automatically adapts to display prohibited content in another configured output like his personal PDA and thereby does not display output when the camera input detects arrival of child into living room during movie display on projector. This envisioned scenario is a simple illustrated advantage of multi-modality in collaboration and several application scenarios can be envisioned with the advent of multi-modal processing for wireless multimedia collaboration.

3 Hybrid Collaboration Framework

3.1 Hybrid Collaboration Architecture

The architecture for multi-modal analysis is based on the hybrid collaboration model shown in Fig. 3. A collaboration agent processes multiple inputs to produce dynamic adaptive context-sensitive output. The basic framework accepts multiple inputs from one or several input devices and the input processing block assigns priorities and rankings for inputs based on pre-defined user settings. The user could pre-assign voice input to have the highest priority and the ranking is based on the strength of the input analyzed by this block. A voice input to “shut down” could be obtained as “xxx down” with “shut” information being lost and thereby dose not resemble a meaningful input to the device. This corresponding input is ranked lowest by the input processing block. The highest ranking-highest priority input is fed to the real time analysis agent and the multimodal analysis agent for hybrid collaboration.

The real-time analysis block performs SISO processing for the best-categorized input and this is then appended to the multimodal output for real

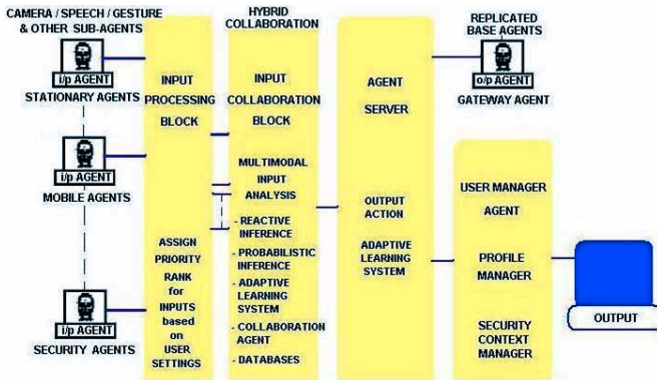


Fig. 3. Multimodal Framework

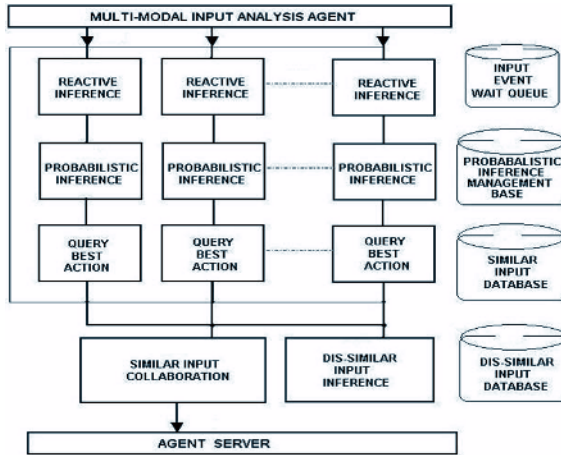


Fig. 4. Hybrid Multimodal Collaboration Architecture

time output action processing at the agent server. The processing at the agent server is deemed real-time since the real-time output from the highest priority highest ranked input waits for a real-time for multimodal action. The multimodality occurs only if the corresponding multimodal action arrives within the real-time threshold defined by the user and formulated by the agent server. Our research calculations have obtained acceptable real-time analysis for 75% similar inputs and 25% dissimilar inputs. The input agents are fed with priorities and rankings from the multimodal analysis agent to the reactive inference agents as a block. The architecture for hybrid collaboration of multi-modal inputs is illustrated in Fig. 4. The reactive inference of the output is then fed to the probabilistic inference agent that analysis the possibility of other inputs and best output action based on previous input blocks and output sets.

The corresponding rankings are fed to the output query agent that saves the similar data and facilitates the best output to the similar input inference and compares with databases for previous input blocks that belong to this output action set. The final output is fed to the agent server that co-works with the user manager agent to provide dynamic adaptive output based on context.

3.2 Learning System Design

The probabilistic inference agent is based on a continuously adapting rule-based learning system. We have adopted the simpler rule-based model and would later incorporate neural agents to enhance this basic model to obtain higher efficiency and accuracy of output action. The pseudo-code for the system is defined as follows:

```

# Pseudo-code for evaluating the correctness
# of probabilistic learning
pi=0; // initialize probabilistic inference count
total=number of inputs; // input count variable
for each input_block {
    for each rule {
        determine if rules are predictable
    }
    predict=predict output action();
    actual_output=query_real_output();
    if actual_output=predict {
        pi++;
    }
}
correctness=pi/total;
// adapt dynamically based on correctness threshold to learn
# end of pseudocode

```

The learning system agent runs on top of the real-time analysis agent, reactive analysis agent and the final collaboration agent to enable the system to dynamically adapt and learn to changing user context scenarios. The learning system operation may have a low correctness initially on initial usage and later adapts to mimic the digital home network scenario as closely as possible with usage.

The usage of adaptive learning system with the hybrid collaboration architecture facilitates the system to adapt to dynamic user requirements based on the concerned user context and other parameters. It facilitates the system to behave as closely to the required adaptation level with usage.

4 Conclusions

The Home Server plays an important role for home networking devices in the digital home. The multitude of home networking devices available to the home user necessitates the usage of the hybrid collaboration architecture for multi-modal processing. The plurality of input agents and security inputs facilitates the mobile agent architecture demonstrated in this paper. The architecture is facilitated with a learning agent to enable adaptive learning and improvisation of the architecture based on user context and usage scenario.

We have identified the basic architectural framework in this paper and in-depth architectural analysis is beyond the scope of this research outcome. Home Server architecture and wireless multimedia collaboration for the digital home with adaptive dynamic multi-modal processing facilitates a plethora of services to facilitate the next generation intelligent home networking devices. The usage of hybrid mobile and stationary agents to facilitate multi-modal output actions

based on context and creating dynamic adaptive learning agents towards a new hybrid architecture envisions a whole new stream of opportunities for the digital home.

Further research in the area needs to simulate the real-world home scenario to the best possible extent and create time-triggered message-triggered objects that provide intelligent object processing information to facilitate real-time multi-modal processing. Digital Home implementation issues concerning bandwidth for wireless multimedia and performance optimization for QoS and security over wireless links are subject to advanced research.

Acknowledgement. This research was supported by University IT Research Center Project and in part by ETRI, Korea.

References

1. Changseok Bae, J.W. Seok, Y.S. Choe, and J.W. Lee: Multimedia Data Processing Elements for Digital TV and Multimedia Services in Home Server platform. *IEEE Transactions on Consumer Electronics*, Vol. 49, No. 1, (2003) 64–70
2. Lee, K.H., Kim, D.H., Kim, J.Y., Sul, D.M., and Ahn, S.H.: Requirements and Referential Software Architecture for Home Server based Inter-Home Multimedia Collaboration Services. *IEEE Transactions on Consumer Electronics*, Vol. 49, No. 1, (2004) 145–150
3. Jagannathan, S., Hosking, A., Welc, A., and Vitek, J.: A Semantic Framework for Design Transactions. *European Symposium on Programming* (2004)
4. Sztipanovits J., Abbott B., Bapty T., Abbott B.: Model-based synthesis of complex embedded systems. *Proceedings of the 1994 Complex Systems Engineering Synthesis and Assessment Technology Workshop (CSESAW'94)*, Washington DC (1994)
5. Kim, D.H., Park, S.M., Kim, J.Y., Sul, D.M., and Lee, K.H.: Collaborative Multimedia Middleware Architecture and Advanced Internet Call Center. *Proc. 15th ICOIN, Betpu, Japan* (2001) 246–250
6. ITU-T Draft Recomm.: T.125 - Multipoint communication service protocol specification
7. Handley, Schulzrinne, Schooler, and Rosenberg: Session Initiation Protocol (SIP) Extension for Instant Messaging. *RFC 3428*, Internet Engineering Task Force, Dec. (2002)
8. Schulzrine, Casner, Frederick, and Jacobson: RTP: A Transport Protocol for Real-Time Applications. *RFC 1889*, Internet Engineering Task Force, Feb. (1996)

E-learning as Computer Games: Designing Immersive and Experiential Learning

Ang Chee Siang and G.S.V. Radha Krishna Rao

Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Selangor Malaysia

Abstract. The article presents the academic views of narrative in the interactive environment particularly in computer games. The relationship between stories and games is examined from the perspective of ludology and narratology in order to understand how computer games work as a medium for storytelling. E-Learning software is thus analyzed as computer games and several issues pertaining to the lethargic stage of educational software development are raised. Then, two versions of e-Learning prototypes are demonstrated and discussed.

1 Introduction

Since the dawn of human history, games have been used in teaching and learning. Board games for example are believed to be the earliest games and they are battle simulations designed to instruct the young. Schools have become the dominant learning environment as learning process is getting complex with the emergence of more complicated knowledge that is too time consuming to learn by playing games. Being born in a world in which pictures flick 25 times a second, the new generation is unable to accommodate in a traditional school learning environment. The advent of computers has opened up a new door for education. However, most educational technologies document a disappointing history. One of the problems is that educational software is rather static and text-heavy. Much of this software turns out to be an electronic version of book, in which no cognitive effort is required to scroll along the boring text. Consequently, it is claimed that the study of multimedia learning could be centered in computer games, because they involve an enormous learning experience [1].

2 Games and Stories

Almost everyone will agree that not all computer games are interactive narrative. In most modern computer games, the players can naturalize their actions as the solving of a familiar type of problems [2]. In *Myst III*, the player needs to track down the villain; in *Super Mario Bros. 3*, the player is trying to save Princess Toadstool. Some games on the other hand, do not feature a concrete setting and can hardly be interpreted as the pursuit of human interests in a concrete situation [3]. This article focuses on games that can be recounted in narrative

discourse and attempts to examine the subtle relationship between games and narratives. It begins by examining the definition of narrative given by Ryan [4]:

“Narrative is defined as a mental image, or cognitive construct, which can be activated by various types of signs. This image consists of a world (setting) populated by intelligent agents (characters). These agents participate in actions and happenings (events, plot), which cause global changes in the narrative world.”

Here, several useful terms are identified: world or setting, character and action. The authors would like to know to what extent these exist in computer games. A game has a spatial representation whether it is real or abstract. Espen Aarseth has claimed that spatiality is the main theme in computer games [5]. *Myst III* for example has a rich description of space represented with high quality pre-rendered 3D images. The players recognize the space immediately after entering the game world, and know how they should act because it is intuitive and resembles a real social setting. *Pong* represents an abstract space, which might not have a referent in the real world. One thing in common is that both spaces operate within a strict set of rules that define the mechanism of the worlds.

Most computer games feature explicit characters that interact with the world or the player. In *Myst III*, the characters are descriptive and have clear characteristics. The players can interact with them as if they are real human although the interaction is limited to several chosen aspects. In *Pong*, even though it does not have an explicit character, the player is playing against an opponent. The character is not presented graphically in the game world, but the existence cannot be overlooked. Finally, all games involve active actions and reactions of the players. In actuality, games can be referred to as a goal-directed and competitive activity (action) conducted within a framework of agreed rules. Games are usually discerned from narrative by the existence of interaction. These actions include not only the action of the player, but also the autonomous actions of the game characters.

It appears that games and narratives are quite similar as computer games use narrative structures to organize their worlds. Nevertheless, games are not a mental image; they are a system that is defined by a set of concrete rules. Within this context, the players can act freely as long as their actions conform to the rules. The chain of these actions can be recounted in narrative discourse and interpreted in the mental image. However, there are sequences of events in games that do not become or form stories (like in *Tetris* for example). Therefore, it is argued that not all games are interactive narratives; rather some games can be interactive narratives and these games can be used as a medium for storytelling.

3 Ludology and Narratology

There is almost no doubt that narratives alone do not make a game. Computer games require a simulation that allows the interaction between the player and the story. Ludology states that a game is organized within its internal structure,

and oriented toward a goal. Unlike narrative readers, game players not only play to know the advancement of a story: their play is centered in a discovery of an open space that invites observation through the duration of temporality [6]. According to Frasca, the structure and the goal of a game are governed by *paidea* rules and *ludus* rules respectively [7]. He identifies two types of game: *ludus*, which refers to the games whose result defines a winner and a loser; and *paidea*, which refers to games that do not. He also recognizes two types of rules: *paidea* rules are rules established in order to play the game, while *ludus* rules are established in order to win or lose the game.

Based on his definitions, a book is also a kind of play, as it has *paidea* rules: the reader must turn the page to read the next part of the text. One can also turn a book into *ludus* by adding *ludus* rules: whoever finds page number 46 the fastest win. In such a way, a book can be a game that tells stories. After experiencing with the pages and the texts, one constructs a story in his or her mind. From here one can see the fundamental difference between the focus of narratology and ludology. Narratologists want to examine how one can tell a good story on this *paidea*, while ludologists choose to study the rules and mechanics of the book or perhaps the interaction between texts and readers, making the page flipping experience or text interaction more interesting.

History had witnessed how people studied the mechanics of book, made it more convenient, portable and able to represent more stories. Even now the printed-paper takes different forms, from the foldable map to the paged dictionary. However, there is not much one can study about the mechanisms of a book as play because of the physical limitation of the medium. People are more interested in the story told in a book. The same concept can be applied to computer games. One can choose to study the story told by the simulation or the simulation itself.

However, books only support the construction of the narrative discourse: the stories are already written into the book. *Paidea* rules of a book define how reader can discover the story. Thus, a book is representing a story. After reading the stories, the readers will construct narrative in their mind. In a computer game, the *paidea* rules define how the narrative space functions and operates. The player interacts with the space, enact stories and construct the narrative. Note that there are also games that represent a story like a book, such as adventure games. Hence, the narrative in games can be categorized into two types: represented and enacted. In most modern computer games, it is usually a mix between the two. The game would represent stories with cut-scenes, and allow the player to construct stories by playing and interacting with the objects in the game world.

4 E-learning as Computer Games

The authors have explained how computer games could be a medium for storytelling. Let us look at the problem statement: why are most traditional e-Learning systems not interesting, or less interesting than commercial computer games? One can even further enquire: is an e-Learning system a game? In order

to answer this question, one needs to look into the definition of *paidea* and *ludus*. E-Learning has *paidea* rules, which are rather symbolic: click the menu buttons and scroll the text with the mouse button, etc. *Ludus* rules are usually stated as the learning objective: to understand the concept of metamorphosis. Like a book, e-Learning software could be *paidea* or *ludus* depending on the existence of an explicit goal. But why is it not as engaging as commercial games? Based on game theories, four reasons are identified:

- The lack of the sense of narrative space
- The lack of story
- The lack of semantic *paidea* rules
- The lack of explicit *ludus* rules

4.1 Narrativity

Narrative interfaces have been used in the game industry since its infancy and have successfully enticed a large portion of computer users for decades. E-Learning software fails to take advantage of this highly effective design. Spatial design is obviously lacking as most interfaces of traditional e-Learning system use the metaphor of a book. The computer screen should not be a representation of a page of book, but a window to a new world. Learners look through the screen like through the window to a new spatial world of knowledge in which the images of real objects act coherently with virtual models [8].

The interface of game is doubled in an interesting way. The first is the interface of the computer: the keyboard and the mouse. An additional interface is the narrative metaphor, which illuminates the narrative space in a new dynamic and interactive medium. The spatial design makes the first interface “disappeared”. The learners are not interacting with the keyboard or the mouse, but the story presented from the computer screen [9]. Another issues pertaining to the spatiality of the software is that most traditional e-Learning system presents learning content linearly, offers textual explanations, and gives a particular spatial organization that does not reflect physical experiences. The learners should not regurgitate the context-free facts; rather they expect to utilize knowledge in a contextually rich situation.

Apart from this, despite the possibility of the simulation to provide both kinds of narrative - enacted and represented - e-Learning does not offer narrativity to its users. In many e-Learning systems, there is hardly any action except for the clicking of the menu buttons, which is hardly conceivable as stories. As Ryan has pointed out, the players do not want to “gather points by hitting moving targets with a cursor controlled by a joystick”; they want to fight terrorists or save the earth from invasion by evil creatures from outer space [4]. It is the same for an e-learning system, which is also a type of game. The learners do not want to click the button to flip through the pages about genetic; they want to defeat the monsters by analyzing and breaking their genetic codes.

According to the motivational heuristic of computer game proposed by Malone [10], narrativity in computer games can provide at least two fun factors:

fantasy and curiosity. Fantasy in computer games is created usually by cladding a narrative layer to the abstract mechanism of the game. E-Learning is unable to create fantasy because of the abstract and symbolic representation and interactions. As a conclusion, e-Learning should be designed with narrativity as follows:

- E-Learning should be designed as a narrative space with story possibilities.
- Narrative world should interact with the learner in a meaningful way.
- Events and actions should be tellable as narratives.

4.2 Paidea and Ludus

Games are more successful in creating emotional reactions through interaction than through storytelling. The internal structure of a game can be characterized by its paidea rules, which can be classified into two types: symbolic and semantic. Briefly, symbolic paidea rules explain the first layer of game interface: the input and output interactions, while semantic paidea rules describe the narrative layer of the interface. Obviously, the paidea rules of most e-Learning system are symbolic, and do not provide semantics to the learners.

The enjoyment of users should not be limited to symbolic paidea rules that define how users interact with the computer devices. Learners should engage in play by observing, hypothesizing, testing and updating the semantic paidea rules of the narrative world. The enjoyment of paidea should lie in the exploration of the virtual world and the discovery of the paidea rules. Unlike Super Mario Bros. 3 where players play and observe the causality of their actions and the behavior of the spatial system, e-Learning systems do not provide such qualities.

The fun factor that paidea rules provide can be associated to Malone's control and curiosity. For paidea rules, the player is curious to test the rules and mechanisms of the game world. Players want to know how the world operates, how the characters or objects act and react. They are curious to test and possibly trying to break the rules to fulfill their incompleteness of the system. Besides, the player must experience feelings of control over actions and environment for the activity to encourage playful, exploratory behavior.

The game designer not only has to design the paidea rules that make the simulation work, but also defines the goal of the game (ludus rules). For a game-based e-Learning system, explicitly stated ludus rules can be important to scaffold learning. Ludus rules also work as the guidance in the virtual world that leads the players to the learning objective. This also matches the task-based learning, while each task is introduced as ludus rules.

As Malone has pointed out, ludus rules are important in order for a computer game to be challenging, especially for naïve learners who cannot identify the goal by themselves. The goal should not be too easy or too difficult to attain for players over a wide range of ability levels. One of the solutions is to create an environment without built-in goals, which is structured so that users are able to generate their own goals. As a conclusion, e-Learning should be designed with play and game activities as follows:

- Paidea rules should simulate and bring about a dynamic narrative space.
- Ludus rules are stated as narrative goals, which will lead the learners to learning objectives.

5 Theory into Practice

In this section, the authors demonstrate two types of e-Learning design by presenting two versions of JapanGO. The first version is to reflect the traditional design of e-Learning, in which information is organized in several chapters. The learners are to read the material from the first chapter to the last, although non-linear reading is possible. They are expected to answer quizzes about what they have learned.

In the second version, the players are invited to play the role of an English teacher in a Japanese high school whose goal is to improve the English standard of the students. In order to communicate with the students and the other teachers, the players need to learn Japanese language. In the school, the players should attend classes and mark homework. They can also visit some places such shopping center, park, etc and talk to people. Figure 1 shows some screenshots of the first and the second version of JapanGO.

Being a typical traditional e-Learning, the first version of JapanGO resembles an electronic book. The computer screen is presenting pages of book to the learners, instead of teleporting them to a virtual environment they subjectively believe is real. Ludus rules are learning objectives instead of fantasy goals. Although certain narrative elements are included, its spatial design is superficial. The player is unable to interact with the character or the environment as they serve only as eye-candy. Besides, there is no action that can be told as narratives. Clicking and scrolling are symbolic actions, which are too abstract to be perceived as narratives. Neither the narrative nor the paidea rules are effective

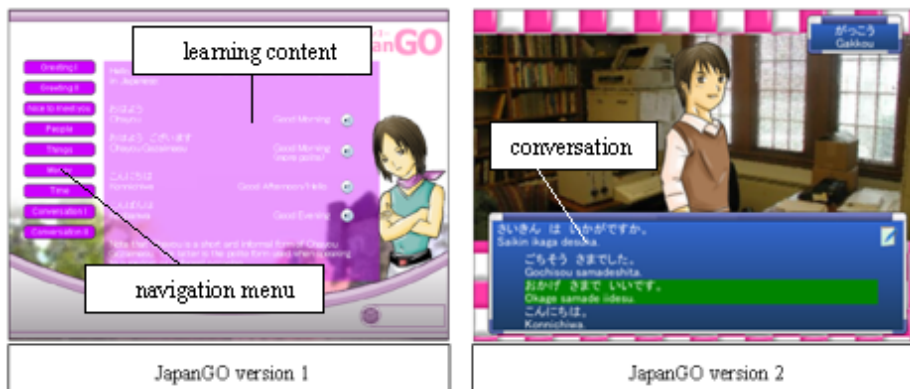


Fig. 1. Two version of JapanGO

to incite curiosity of the learner. Everything is revealed to the player immediately after the program is run. All semantic paidea rules are identified almost instantly: click and the learning content shows. User control is very limited. Although the players do have control over how the program operates, they are unable to control the learning content, which is presented in a fairly fixed way. Since the learners are just interacting with abstract symbols operating within paidea rules, it fails to create fantasy. Lastly, it is not challenging to operate the program. No effort is needed to understand the paidea rules.

Contrary to the first version, the second version offers learners the exploration and discovery of semantics. Learners plan for strategies in order to achieve the ludus rules. They also create their own ludus rules, for example to reduce the homework load level or to increase the student's obedience level. This version has rich spatiality and stories. Instead of being merely a graphical representation, the characters can interact with the player and this has an impact on the game space. The symbolic interaction is clad with another narrative layer and translated into narrative actions such as to sleep, or to read.

The learners are curious to know both the narrative and the mechanism of the virtual world. They are given more access to the places and to meet more characters in the game with the advancement of their Japanese language ability. They may also want to understand the game mechanisms: "what if the student obedience level is high?" The learners have not only control over the program, but also the learning content. Instead of presenting all information at once, the content is revealed gradually depending on the level of the learners. Fantasy is intrinsic as the fantasy is also dependent on the skill (Japanese language); the skill is parts of the fantasy world. Finally, the explicit ludus rules provide challenging tasks to the learners. The learning of Japanese language is important to achieve the goal.

Table 1. Expected results and the rationale

Expected Results	Rationale
Challenge	It provides challenging environment with explicit ludus rules.
Fantasy	Narrative layer creates fantasy by giving meaning to the abstract mechanisms of the game world.
Curiosity	Both narratives and paidea rules incite the learner's curiosity.
Control	Unlike traditional narrative, the game is simulated by paidea rules that provides a higher degree of control over both the program and the learning material.
Information retention	Learning material which is integrated into a narrative context is easier to be remembered because the information is also connected in a network-like structure in our long term memory.
Knowledge transfer	Since game design provides a context-based learning environment, it is expected that the learned knowledge can be transfer to a real situation more easily.

5.1 Expected Results

It is maintained that game theories provide a better framework for designing e-Learning, making the experience of learning more immersive and engaging. The game-based e-Learning design is expected to be better than the traditional one from the following perspectives as shown in Table 1.

6 Summary

In this work, the authors are more interested in narrative games especially when learning is in question, as it is believed that narrative games are able to trigger the emotion as the heart of the game is its dramatic force; rather than a lecture, the player is compelled by an emotional logic. Future research is to be carried out in order to verify the expected result.

References

1. Ang Chee Siang, GSV Radha Krishna Rao: Theories of learning: A Computer Game Perspective, IEEE Fifth International Symposium on Multimedia Software Engineering (2003)
2. Janet Murray: Hamlet on the Holodeck. The Future of Narrative in Cyberspace. New York: The Free Press (1997)
3. Marie-Laure Ryan: Beyond Myth and Metaphor – The Case of Narrative in Digital Media, the international journal of computer game research volume 1, issue 1 July (2001)
4. Marie-Laure Ryan: Immersion vs interactivity: Virtual reality and literary theory, Postmodern Culture (1994) available at <http://muse.jhu.edu/journals/postmodernculture/v005/5.1ryan.html>
5. Espen Aarseth : Allegories of Space: The question of Spatiality in Computer Games (1998) available at <http://www.hf.uib.no/hi/espen/papers/space/>
6. Bo Kampmann Walther: Playing and Gaming Reflections and Classifications, the international journal of computer game research volume 3, issue 1, May (2003)
7. Frasca Gonzalo: Video Games of the Oppressed: Video Games as a Means for Critical Thinking and Debate, Georgia Institute of Technology (2001)
8. M. N. Morozov, A. I. Markov: How to make courseware for schools interesting: new metaphors in educational multimedia, International Workshop on Advanced Learning Technologies. Advanced Learning Technology: Design and Development Issues, IEEE Computer Society (2000)
9. Lanier Jaron, Frank Biocca: An Insider's View of the Future of Virtual Reality, Journal of Communications 42.4 (1992): 150-172.
10. Thomas W. Malone: What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games, Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems (1980)

Event-Based Surveillance System for Efficient Monitoring

Do Joon Jung¹, Se Hyun Park², and Hang Joon Kim¹

¹ Department of Computer Engineering, Kyungpook National Univ., Korea
{djjung,hjkim}@ailab.knu.ac.kr

² School of Computer and Communication Engineering, Daegu Univ., Korea
sehyun@daegu.ac.kr

Abstract. In this paper, we propose an event-based surveillance system which detects the events in the image sequence obtained from each camera and the detected events images serves to the operators via internet. In the proposed system, we decided the thirteen events which are helpful to the operator such as enter, leave, picks up briefcase, use computer in the room environment. We detected the events through variations which are inside of the object and position relation between objects. In the experiment, the system had successfully detected these events in several images sequence, the success rate of event detection was 88% on average.

Keywords: Event Detection, Surveillance System, Efficient Monitoring.

1 Introduction

Increased personnel expenses and reduced costs for surveillance system make a situation where operators simultaneously monitor multiple environments. But it is widely known that, if the operator is exposed to this type of work for several hours, his attention decreases, thus the probability of missing dangerous situations increase [1].

G.L.Foresti describes a visual surveillance system [2] for remote monitoring of unattended outdoor environments. The system is able to detect, localize, track, and classify multiple objects moving in a surveillance area. But the system has limitations that are mainly due to low illumination and the presence of shadows which can mislead the system and force it to erroneously generate alarms.

A.C.M.Fong introduces a web-based surveillance system [3] that allows remote monitoring via internet. It is a useful system, when you need a remote surveillance as far as long distance.

The system which introduced by G.L.Foresti has a shortcoming which is the operator must take a position nearer to the installed camera than other web-based surveillance system. The system introduced by A.C.M.Fong is more efficient than G.L.Foresti's system in terms of transmission area, but the operator must always attend the monitor for monitoring. Because of the monitoring over long periods of time, the attention of the operator is easily decreased and the missing rate is increased. If some observed events in the surveillance area can be

provided to the operator, the operator's effort would be decreased and efficient monitoring will be possible. Therefore, we propose an event-based surveillance system which detects the events in the image sequence obtained from several cameras and the detected event images serves to the operator via internet.

2 Proposed System

2.1 Overview

Figure 1 shows the configuration of the proposed system. The proposed system consists of four digital cameras and six computers. Four computers are used as an event detector, other one computer is used as a monitoring server, and another one computer is used as a monitoring client. Each digital camera is connected to a computer which is responsible for event detection of the video captured from the digital camera. Each computer is connected to the web server through a local area network (LAN).

If an event detector detects any event, it will send the event images to the web server. The event images in the web server can be requested by the monitoring client over the internet. Each camera is located at a different location for wide area surveillance and detects an event in a scene.

The proposed surveillance system consists of three parts. These are an event detector, a monitoring server, and a monitoring client. The event detector is a remote station located on the site under surveillance. Its role is video capture, event detection of captured video, and transmission of captured video and detected event images to the monitoring server. The monitoring server's role is database management and monitoring client's request management. The monitoring server receives the captured video and event images, and then saves the received video and event images in the database. The monitoring client's role is

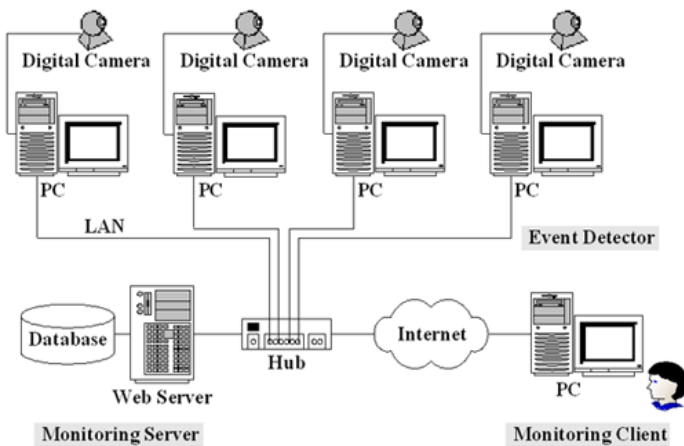


Fig. 1. Configuration of the proposed system.

requesting the captured video and the event images from the monitoring server. In the monitoring client, there is a graphical user interface (GUI) and it is performed through a web browser. Each event image in the monitoring server is requested by the monitoring client according to the user action and then it is displayed in the web browser.

2.2 Event Detection

In the proposed system, the event is detected through variations which are inside of the object and position relation between objects. We extracted features which are skin region and object variation using the three vision techniques. Three vision techniques are the skin region detection, tracking, and change detection.

In order to reduce the computation time for event detection and easily detects the event, we assume that the interesting place will be known. Figure 2 shows the scene configurations for event detection. In the figure 2, the bounding boxes are the interesting places and its initial position must be defined by operator. The red bounding boxes are the place where the person will appear. The green bounding boxes are places which are around the interesting object. Skin region detection is only performed in the red bounding box and change detection is performed in the green bounding box and whole image.

For event detection, we must decide the start and end instant of event, because the event is based on a time interval, not on a time point. Therefore, we defined the start instant of all events when change is detected in the whole image. Each event is detected with event detection conditions. Table 1 shows the event detection conditions.

In the proposed system, the event detection method consists of feature extraction module and event detection module. Figure 3 shows the architecture of

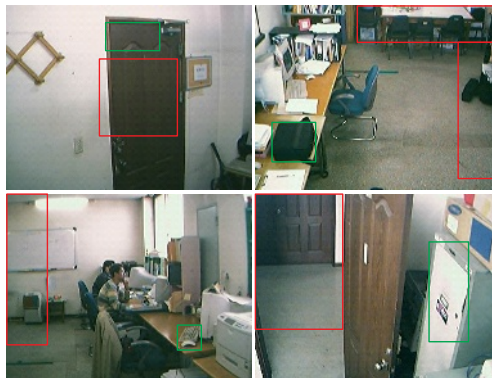


Fig. 2. The scenes configuration. The first scene on the left is set up to show a person entering or leaving. The second scene on the right is setup to show a picking or putting down or leaving with briefcase. The third scene under the left is setup to show a person using a computer. The final scene is setup to show an opening of a cabinet.

Table 1. Event detection conditions

Event	Event Detection Conditions
Enter/Leave	Intensity variation is detected at the door and dominant skin region has appeared / disappeared.
Sit/Stand	The y-position of “tracked skin region” is decreased / increased significantly.
Near to something	The position of the “tracked skin region” move to the briefcase, cabinet or computer.
Pick up briefcase	The position of “tracked briefcase region” is moved significantly.
Put down briefcase	The position variation of the “moved briefcase region” has stopped.
Leave with briefcase	The event “pick up briefcase” has occurred and “tracked skin region” disappeared before the event “put down” has occurred.
Use computer	The position of the “tracked skin region” move to the computer and intensity variation is detected at the mouse or keyboard.
Open a cabinet	The position of the “tracked skin region” move to the cabinet and intensity variation is detected at the door of the cabinet.
Close a cabinet	The event “open a cabinet” is detected and intensity variation is detected at the door of the cabinet.

the event detection method. In the feature extraction module, three features are extracted. These are a position of skin region, a position of object region, and intensity variation of object region. Initial position of object is given by user, so it is only tracked by tracking technique. The intensity variation of object is detected through change detection technique. In the event detection module, the thirteen events are detected and detected event images are stored in the each event category with detected time. Event detection module consists of thirteen event detection sub modules. Each event detection sub module is responsible for an event detection. And each sub module detect events according to the event detection conditions which are described above in the table 1.

2.3 Vision Techniques

The event detection performance depends on the accuracy of the vision techniques. Therefore, our vision techniques are a modified version of the techniques in the previously introduced vision techniques which had good performance.

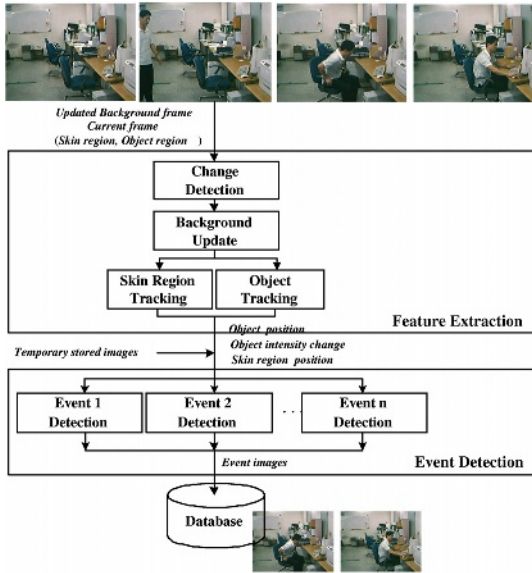


Fig. 3. Architecture of the event detection method.

Skin Region Detection. The color distribution of human faces is clustered in a small area of the chromatic color space [4]. When the chromatic r and chromatic g of skin pixels from face patch are plotted in $CrCg$ -space, skin color occupy with elliptical shape. Figure 4 (a) are the sample face patches and (b) shows the skin locus of a USB Smile-cam camera used in the proposed surveillance system. A simple membership function to the skin locus is a pair of quadratic functions defining the upper and lower bound of the cluster [5].

In the experiment, the shape of skin cluster is similar to an ellipse. Therefore, we decide the membership function to the skin locus is an elliptical function as follows:

$$S = \begin{cases} 1, & \frac{(x' - c_x)^2}{a^2} + \frac{(y' - c_y)^2}{b^2} < 1, \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \quad (2)$$

where $C_x = 0.347$, $C_y = 0.372$, $a = 0.074$, $b = 0.022$, $\cos \theta = 0.840$, and $\sin \theta = 0.542$ are computed from the skin cluster in the $CrCg$ -space.

Tracking. The tracking techniques of the proposed system uses a method reported in [6], which is based on statistical color modeling and the deformable template. First, for each pixel in the current frame, the method calculates the

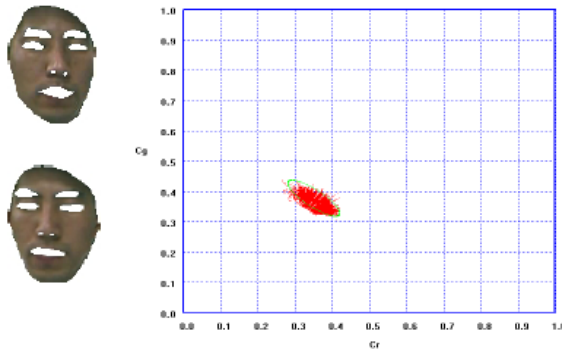


Fig. 4. (a) The sample face patches, (b) Skin locus of Smile-camera in $CrCg$ -space.

probabilities of each pixel belonging to each of the two classes: the face class and the non-face class. Then the method uses a deformable template to group the pixels more likely to belong to the face class. The method deforms the template so that it includes as many face pixels as possible and at the same time includes as few non-face pixels as possible. We applied this method for skin region tracking and object region tracking.

Change Detection. We limited the start of event as on a time point when change is detected. For the decision of the start of an event, we used the change detection technique. Moreover, we applied the change detection technique on the region of interest (ROI) for checking the variation of a fixed object. We used change detection technique which is reported in [7] because we need a method which has relatively low level computation cost and camera view was fixed. We create the background model BG_i by using the first frame F_1 . And then we detect the foreground region FG_i through background subtraction. The foreground region is defined as follows:

$$FG_t(x, y) = \begin{cases} 1, & |BG_t(x, y) - F_t(x, y)| \geq \theta, \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where $BG_t(x, y)$ and $F_t(x, y)$ are the intensity values at the background model and the current frame, respectively. $FG_t(x, y)$ is a binary image that indicates the foreground and background. The threshold value θ is 30. We update the background model using current frame F_i and foreground region FG_i . We generate the dilated foreground region FG_{di} which will be used for background updating. We only update the background on the dilated foreground region where $FG_{di} \neq 0$. The skin region detection is performed only in the foreground region because to prevent that the skin like pixel is detected in the background.



Fig. 5. The GUI of the proposed system.

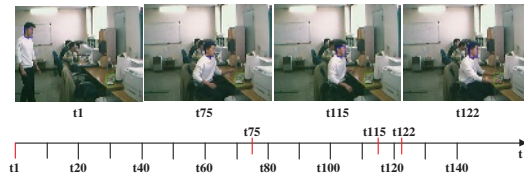


Fig. 6. Event detection result (use computer).

3 Experimental Results

In the proposed surveillance system, the operator can be watching the surveillance area through the internet browser. Figure 5 shows the graphical user interface (GUI) of the proposed system. The system allows the real time monitoring and retrieving the event images. Moreover, the system has the log-in function for only service to the authorized operator. The event images are added up to the event list in the real time.

Figure 6 shows the example of event detection results. In the figure 6, the y-position of tracked skin region is decreased significantly so the sit event is detected at t_{75} , the position of tracked skin region stopped near computer so the near computer event is detected at t_{115} , and change is detected in the mouse and keyboard so the use computer event is detected at t_{122} .

The system for detecting events has been tested in the room environment (the laboratory). In the test, the video sequences were acquired with a frame rate of 15 frame/sec and the size of each frame is 320×240 pixels. For the test, we considered about 1300 events equally distributed on the thirteen event classes. The success rate of event detection was 88% on average.

4 Conclusions

In this paper, we proposed an event-based surveillance system which detects the events in the image sequence obtained from each camera and the detected events images serves to the operators via internet. In the proposed system, we decided the thirteen events which are helpful to the operator in the room environment. We detected the events through variations which are inside of the object and position relation between objects. We extracted some features which are skin position, object position and object variation using the three vision techniques for event detection. In the experiment, the system had successfully detected these events in several image sequences, the success rate of event detection was 88% on average. Therefore, the proposed system decreases the probability of missing dangerous situations and the system will help to the remote operator who is exposed to monitoring over long periods of time.

Acknowledgements. This research was supported by Kyungpook National University Research Fund, 2002

References

1. Stringa, E.S., Regazzoni, C.S.: Real-Time Video-Shot Detection for Scene Surveillance Applications. *IEEE Transactions on Image Processing*, Vol. 9 (2000) 69–79
2. Foresti, G.L.: A Real-time system for Video Surveillance of Unattended Outdoor Environments. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 8 (1998) 697–704
3. Fong, A.C.M., Hui, S.C.: Web-based intelligent surveillance system for detection of criminal activities. *Computing & Control Engineering Journal*, Vol. 12 (2001) 263–270
4. Yang, J., Waibel, A.: A Real-Time Face Tracker. *IEEE Workshop on Application of Computer Vision*, (1996) 142–147
5. Soriano, M., Martinkauppi, B., Huovinen, S., Laaksonen, M.: Adaptive skin color modeling using the skin locus for selecting training pixels. *Pattern Recognition*, Vol. 36 (2003) 681–690
6. Huang, F.J., Chen, T.: Tracking of Multiple Faces for Human-Computer Interfaces and Virtual Environments. *IEEE International Conferences on Multimedia and Expo*, Vol. 3 (2000) 1563–1566
7. Chen, F.S., Fu, C.M. Huang, C.L.: Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, Vol. 21 (2003) 745–758

A Collaborative Multimedia Authoring System Based on the Conceptual Temporal Relations

Mee Young Sung

Department of Computer Science & Engineering, University of Incheon
177 Dohwadong, Namgu, 402-749 Incheon, South Korea
mysung@incheon.ac.kr

Abstract. We developed a SMIL-based collaborative multimedia authoring tool supporting a mechanism for conceptually representing the temporal relations between different media. Among the many editors that make up our system, the temporal relation editor provides users with an intuitive mechanism for representing the conceptual flow of a presentation by simple and direct graphical manipulations. Our system proposes TRN (Temporal Relation Network) as its internal multimedia presentation representation. The TRN corresponds exactly to the conceptual temporal structure of the multimedia presentation. A TRN is composed of media objects, delay objects and a set of temporal relationships among objects. A media object is associated with a duration. A parallel relationship found in a TRN can be collapsed into a single par (parallel) synchronization block. This collapsible synchronization block facilitates the determination of the playing time of each component and can be the basic unit for reusability of already prepared blocks of presentation code. In addition, our system allows users in different places to design together a multimedia presentation collaboratively in reviewing the same presentation at the same time.

Keywords: Collaborative authoring, Multimedia authoring, SMIL (Synchronized Multimedia Integration Language), Temporal relation representation, Synchronization.

1 Introduction

The key to authoring a presentation lies in the composition of temporal relationships between objects. Conceptually, the temporal relationships between two media can be classified into one of seven possibilities. They are ‘before’, ‘meets’, ‘overlap’, ‘during’, ‘starts’, ‘finishes’, and ‘equals’ [1][2]. Every temporal relationship can be described using one of these seven relations. Spatial relationships can be described by specifying sub-regions within the total presentation region that correspond to each object.

The goal of this study is to develop an easy and efficient multimedia authoring environment where users can create a multimedia presentation in a simple and intuitive manner. Toward this goal, we provide users with the capability to

edit temporal relationships between media objects at the conceptual level: for example, presenting object A before B, presenting object A during B, etc. We also want to allow users to create multimedia content without manually specifying the playing time (e.g. a specific start time and duration) for each media. Instead, our authoring system automatically calculates the playing time and then generates proper start times and durations for each object. In the traditional scaled timeline approach, users can directly view and control the structure of the content; however, the representation is fixed, and the operations are manual. Our goal was to develop a good tool for generating the presentation schedules conceptually without considering the details, and the system can automatically detail the properties of the media. Using our system, users can focus on the creative aspects of their design, and not worry about manual specification of timing details for each object.

We developed a multimedia authoring system based on the SMIL (Synchronized Multimedia Integration Language) [3][4][5][6] 1.0 Recommendation[3] and SMIL 2.0 Recommendation [4][7][8]. The existing SMIL authoring tools provide basic user interfaces such as the scaled timeline-based user interfaces (representing media objects as different bars arranged in multiple layers on the scaled timeline) or textual tag editing user interfaces for authoring. What distinguishes our system is that it provides a simple and intuitive editing mechanism for creating conceptual flows of a presentation, in addition to the basic timeline-based interface.

In this paper, we present the design and implementation of our multimedia presentation authoring system which provides a mechanism for conceptually representing the temporal relations of different media. We will examine our mechanism for representing conceptual temporal relationships in the following section. In section 3, we will investigate the Temporal Relation Network (TRN) upon which our model is based. We will discuss our algorithm for automatically generating a TRN from the DOM in section 4. In section 5, the implementation of our collaborative authoring is presented. Finally, the last section will provide conclusions and some future work.

2 Representation of Conceptual Temporal Relations

A main focus in authoring a multimedia presentation is the design of the temporal behaviors for the components that make up the presentation. Our system is designed to allow users to specify temporal behaviors of media objects at the conceptual level. This section describes our model for representing conceptual temporal relations.

Our system's multimedia representation is based on Allen's temporal intervals [1]. Allen distinguished thirteen different time intervals between two objects. They can be reduced into seven temporal relationships such as 'before', 'meets', 'overlap', 'during', 'starts', 'finishes', and 'equals' by removing the relationships in inverse order. The graphical representations of the seven conceptual temporal relations of our system are summarized in Figure 1. The graphical representa-

tions shown in Figure 1 correspond exactly with the internal representation of each corresponding temporal relationship. Note that we represent the parallel relationships such as overlaps, during, starts, and finishes) by adding dummy delay objects to make the 'equal' relationship. As shown in Figure 1, all five parallel relations can be generalized as the 'equal' relation by inserting some delay objects when they are needed. As a consequence, any parallel relation can be collapsed into a single object. We simplify a group of networked icons as a synchronization block object. This block object can be opened to show the details or collapsed to a single block icon as shown in Figure 2. This mechanism is proposed to simplify the representation of complicated parallel relationships within SMIL content.

3 TRN (Temporal Relation Network)

The authoring process is composed of a series of user interactions for editing a multimedia presentation. An interactive authoring system should process each user interaction immediately and return appropriate feedback. Supporting an interactive authoring environment requires consistent internal maintenance of the state of the presentation [9][10][11][12]. Some existing studies on the internal representation of multimedia applications include: OCPN (Object Composition Petri Nets) [2], DTPN (Dynamic Timed Petri Nets) [9], XOCPN (eXtended OCPN) [10][11], etc. DTPN and XOCPN are variants of OCPN. In these systems, the internal multimedia representations are based on the Petri Net, and the interface is a scaled timeline-based UI. Our system uses TRN (Temporal Relation Network) as its internal representation of a multimedia presentation in order to unify the internal representation and the external user interface with respect to temporal relationships. TRN is a directed and weighted graph. Details are described in the reference [13].

As mentioned before, our system is based on SMIL, whose grammatical structure is the same as that of XML (eXtensible Markup Language). SMIL documents, like any other XML-based document, can be described as a Document Object Model (DOM). The objective of DOM is to provide a standardized interface for accessing XML-based documents (such as XML, SMIL, WML; Wireless Markup Language, SVG; Scalable Vector Graphic, etc.) in diverse computing environments. DOM specifies how to describe the logical structure of XML documents and the details of the components that they contain. DOM describes the logical relationships of document components in a tree structure; however, DOM cannot effectively describe the temporal relationships of components in its simple tree structure. Therefore, we need another mechanism to describe the temporal relationships among different media.

Our system represents internally, as well as graphically, the temporal relationships of a presentation in the TRN graph structure. TRN represents temporal relationships among objects and playing times of objects using information from all objects included in document. If the presentation is new, a new TRN is created. If the presentation already exists, the corresponding DOM structure of the

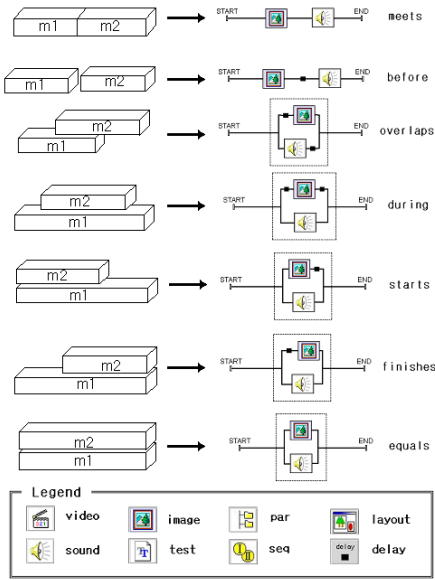


Fig. 1. Representation of temporal relations

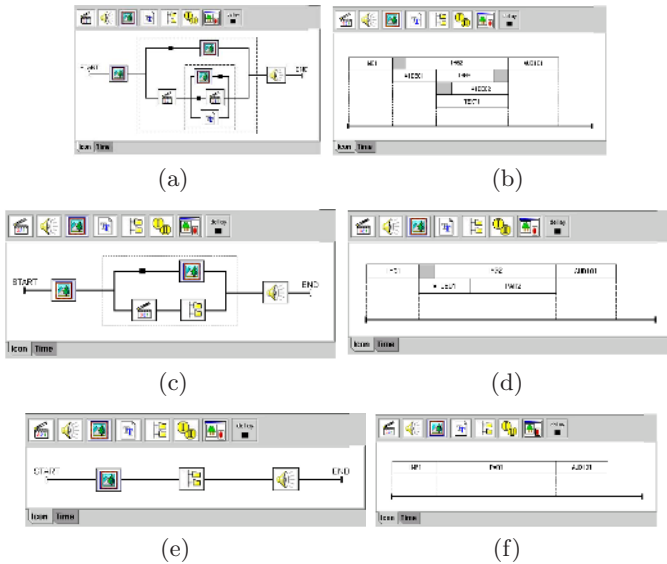


Fig. 2. An example of collapsing some nested parallel synchronization blocks and the corresponding timeline representation: (a), (c), (e) are conceptual representations and (b), (d), (f) are the corresponding timeline representations

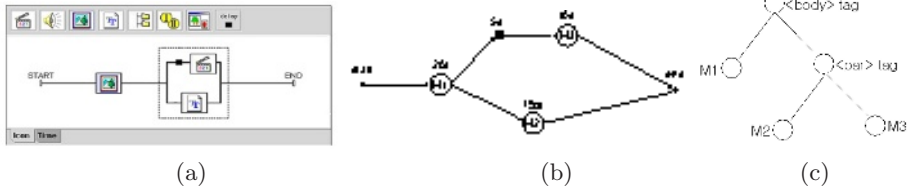


Fig. 3. (a) An example of graphical representation of a multimedia presentation, (b) TRN generated from the DOM information, (c) DOM structure generated from the corresponding TRN.

presentation is reconstructed from its SMIL codes and automatically transformed into the internal TRN structure. As authoring is performed, the underlying TRN must be dynamically changed. After the authoring is finished, a DOM structure can be generated from the internal TRN structure. Our system generates SMIL documents through the interaction between TRN and DOM.

Figure 3(b) illustrates an example TRN that is created as a user authors the presentation shown in Figure 3(a). A node contains more than one outgoing arrow to indicate that the following objects will be played synchronously.

Our system generates the DOM structure in Figure 3(c) and the following SMIL codes from the information of TRN in Figure 3(b).

```
<seq>
  <video id=M1 src=media1.mpg dur=20s/>
  <par>
    <audio id=M3 src=audio.wav begin=5s dur=10s/>
    <img id=M2 src=image.jpg dur=15s/>
  </par>
</seq>
```

Note that there is a direct correspondence between the internal TRN representation and the graphical representation used for authoring a presentation. Therefore, for efficiency, we can collapse parallel relationships found in a TRN into a single node just as we collapsed parallel objects into a single synchronization block as described in section 2.2. This simplification is made through maximized use of the temporal relation ‘equals’. It is important to note that this simplification impacts the algorithms used to calculate the actual values of ‘dur’, ‘begin’, and ‘end’ tags when the SMIL code is generated. Using collapsible parallel objects, our system can easily determine the playing time of each component by first considering the group of parallel relations as a single object. This simplification also impacts the algorithms used to implement a SMIL player, because the SMIL player uses the same TRN when it actually schedules the presentation. In addition, this synchronization block representation will be the basic unit for reusability of SMIL code.

4 Automatic Generation of a TRN from the DOM

The algorithm for automatically generating a TRN from the DOM primarily consists of three modules. They are *build_TRN()*, *insertSeqNode()*, and *insertPar()*.

The *build_TRN()* function actually takes charge of traversing over all the nodes of the document structure. Each component module in this algorithm includes all of the methods required to allow direct or sequential traversal of the document structure, e.g. *getNextSibling()*, *getChildNode()*, *getChildNodes()*, *getParentNode()*, etc. The *insertSeqNode()* routine creates a media node and inserts it into the TRN using the attributes specified as arguments. An additional delay object is automatically created and inserted into the TRN if it is needed. In the *insertSeqNode()* module, the temporal relation ‘meets’ or ‘before’ can be determined by whether or not a delay object exists between the current object and the preceding object. The module *insertPar()* for handling parallel relationships (such as ‘equals’, ‘starts’, ‘finishes’, ‘during’, and ‘overlaps’) of objects. Any parallel relation can be collapsed into a single object. We call a group of networked objects in parallel relationships as a parallel block. The *insertPar()* performs the required tasks as follows:

1. Determine the number of child nodes of the parallel block.
2. Calculate the total playing time of a parallel block.
3. Determine the temporal relationships between each child object and the parallel block.
4. If there are only two child nodes, insert these objects and determine the temporal relationship from the total playing time and the attributes of each child object. Then the routine is terminated.

Note that *insertPar()* is a recursive algorithm for inserting inner parallel blocks inside a parallel block. We also note that Our algorithm should take time $O(MN)$, where M is the number of attributes and N is the number of nodes in the TRN. The algorithm traverses the document tree in $O(N)$ time and each iteration of traversal invokes *insertObject()* which takes time $O(M)$. In practice, the best algorithm for traversing a tree takes time $O(N)$. The data structure itself should store in $O(N)$ space.

5 Implementation of Collaborative Authoring

Our authoring system allows a group of users working at different machines to work on the same multimedia presentation and to communicate in real time. The collaboration manager of our system takes charge of the communications of all events generated by users. Each authoring system at different places can be a server as well as a client of a collaboration group at the same time. A server generates itself as the first client of the collaboration group. Any client can connect to the server using TCP (Transmission Control Protocol) and generates packets corresponding to the content that is created as users edit

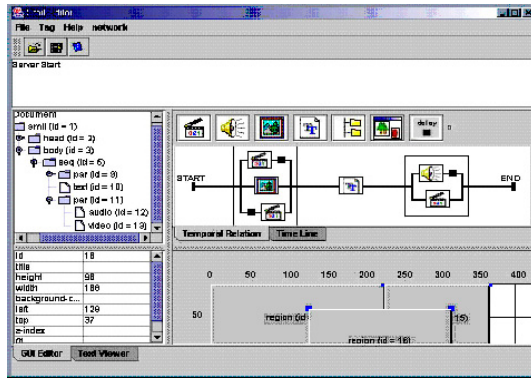


Fig. 4. A screen capture of our authoring system

the presentation. It also receives packets from the server, analyses the packets, and invokes appropriate events or modules. Once a client connects to a server, the server updates the list of groups and initializes the new client by sending a group of objects that have been authored up until that time to the new client. After then, the server multicasts any messages passed to it and the client processes and visualizes any received messages. This mechanism is a variation of a client-server mechanism which can provide better network performance and better portability of the system.

In any collaborative computing environment, multiple users or processes can access a shared object concurrently. In this situation, an inconsistency of shared data might occur therefore a concurrency control is required. We implemented some ideas for efficient concurrency control in our system. They are mainly based on user awareness, multiple versions, and access permissions of shared objects. Details of our concurrency control mechanism are described in the reference [13].

Figure 4 presents a screen capture of our system. Exactly the same images are shown at each user's screen.

6 Conclusion

We developed a SMIL-based collaborative authoring system which allows users to edit the temporal relations among media conceptually by simple and intuitive graphical manipulations. The system editors exchange information through the SOM (SMIL Object Manager) and together form an easy and efficient editing environment. Our authoring system creates and modifies a multimedia presentation using a Temporal Relation Network (TRN) which corresponds exactly to the structure seen in the graphical representation of the presentation. One advantage of our system is that multimedia authors need not specify all of the painstaking details about start times and duration information when creating a presentation. The TRN representation provides an efficient means for the system to automatically fill in the necessary timing details. This frees multimedia authors to focus

instead on the creative aspects of the presentation. Another advantage is the use of collapsible synchronization blocks to efficiently represent and specify simultaneous presentation of parallel media. The collapsible synchronization blocks used by our system also provide a means for portable and reusable SMIL code blocks. Note that in addition to their use in SMIL authoring systems, our system editor modules can be used separately in various kinds of multimedia presentation authoring systems, such as XMT (eXtensible MPEG-4 Textual format)-based authoring tool. Because our system makes use of the standard DOM structure, its components can easily be applied by any multimedia system which also uses the DOM structure for its internal document representation.

Acknowledgement. This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Multimedia Research Center at the University of Incheon.

References

1. James F., Allen: Maintaining Knowledge about Temporal Intervals. *Communications of the ACM* (November 1983) 832–843
2. Thomas D. C. Little, Arif Ghafoor: Spatio-Temporal Composition of Distributed Multimedia Objects for Value-Added Networks. *IEEE Computer* **24**(10) (October 1991) 42–50
3. W3C, Synchronized Multimedia Integration Language (SMIL) 1.0 Specification: W3C Recommendation 15-June-1998, <http://www.w3.org/TR/REC-smil/> (1998)
4. W3C, Synchronized Multimedia Integration Language (SMIL 2.0): W3C Recommendation 07 August 2001, <http://www.w3.org/TR/smil20/> (2001)
5. W3C, W3C Issues SMIL as a Proposed Recommendation, <http://www.w3.org/Press/1998/SMIL-PR> (1998)
6. W3C, Synchronized Multimedia, <http://www.w3c.org/AudioVideo>
7. Dick C.A. Bulterman: SMIL 2.0 Part1: Overview, Concepts and Structure. *IEEE Multimedia* **8**(4) (October–December 2002) 82–88
8. Dick C.A. Bulterman: SMIL 2.0 Part2: Examples and Comparisons. *IEEE Multimedia* **9**(1) (January–March 2002) 74–84
9. Mitsutoshi Iino, Young Francis Day, and Arif Ghafoor: An Object-Oriented Model for Spatio-Temporal Synchronization of Multimedia Information. *IEEE International Conference on Multimedia Computing and Systems*, May 14-19, 1994, Boston, Massachusetts (1994) 110–119
10. B. Prabhakaran and S. V. Raghavan: Synchronization Models For Multimedia Presentation With User Participation. *Proceedings on ACM Multimedia 93*, 1-6 August 1993, Anaheim, California (1993) 157–166
11. Naveed U. Qazi, Miae Woo, and Arif Ghafoor, "A Synchronization and Communication Model for Distributed Multimedia Objects," *Proceedings on ACM Multimedia 93*, 1-6 August 1993, Anaheim, California (1993) 147–155
12. Junewha Song, G. Ramalingam, Raymond Miller, Byoung-Kee Yi, Interactive authoring of multimedia documents in a constraint-based authoring system. *Multimedia Systems* **7** Springer-Verlag (1999) 424–437
13. M.Y. Sung, D.H. Lee: A Collaborative Authoring System for Multimedia Presentation. *Proceedings on The 2004 International Conference on Communications (ICC 2004)*, Paris, France, June 20-24, 2004, MM04-2 (2004) 1396–1400

Multimedia Integration for Cooking Video Indexing

Reiko Hamada¹, Koichi Miura¹, Ichiro Ide², Shin'ichi Satoh³, Shuichi Sakai¹,
and Hidehiko Tanaka⁴

¹ The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{reiko|miura|sakai}@mt1.t.u-tokyo.ac.jp

² Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
ide@is.nagoya-u.ac.jp

³ National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
satoh@nii.ac.jp

⁴ Institute of Information and Security,
2-14-1 Tsuruya-cho, Kanagawa-ku, Yokohama, 221-0835, Japan
tanaka@iisec.ac.jp

Abstract. We have been working on the integration of video with supplementary documents, such as cooking programs. We propose an integration system that performs semantic segmentations of video and text and associates them together. This association is realized using the ordinal restriction of the recipe, cooccurrences of words in the text and the audio in the video, and the relation between the background in a video and words which describe the situation in a text. In this paper, we will introduce the result of an evaluation experiment and show the effectiveness of the proposed integration method. Through our method, many applications should become possible, such as a cooking navigation software.

Keywords: Indexing, Cooking Videos, Association of Video and Text.

1 Introduction

Reflecting the increasing importance of handling multimedia data, many studies are made on indexing to TV broadcast video. Multimedia data consist of image, audio and text, where various studies on analysis of each individual medium have been made. Especially, image processing has been the main medium to handle multimedia data for a long time. But recently, it has started to be considered that image processing alone is insufficient for thorough understanding of multimedia data. From the 1990s, integrated processing that supplements the incompleteness of information from each medium has become a trend [4].

Following this trend, we are trying to integrate TV programs with related documents, taking advantage of the relative easiness of extracting semantic structures from text media. Among various programs, educational programs are considered as appropriate sources, since (1) supplementary documents are available,

and (2) the video contains a lot of implicit information that integration could be helpful to thorough understanding of both media. In addition, the demand for cooking videos and their applications are high, since cooking is a daily and important activity.

Therefore we propose an integration system that performs semantic segmentations of video and text and associate them together. This association is realized using the ordinal restriction from the recipe, cooccurrence of words in the text and the audio in the video, and the relation between the background of the video and words which describe the situation in the text.

In this paper, we will introduce the result of an evaluation experiment and show the effectiveness of the proposed integration method. Through our method, many applications should become possible, such as a cooking navigation software.

2 System Overview

In our system, multimedia data is created from a cooking video by matching it to a corresponding part in a recipe text. An example of the matching is shown in Fig. 1.

As shown in Fig. 1, in cooking programs, the order of steps often differs between a video and a textbook. In that case automatic association of a cooking video and its text recipe is a difficult task. In this paper, a solution which combines information derived from multiple media is proposed. The overview of the integration method is shown in Fig. 2.

First, text segmentation is performed by extracting important words from a text using a domain-specific dictionary. The text is divided into semantic segments. This segment is called a “text block”. Finally the ordinal structure between text blocks is analyzed.

Meanwhile, shot detection, categorization, and background classification are applied to a video corresponding to the text. Next, shots with a same background type are clustered. This shot cluster is called a “video scene” in this paper.

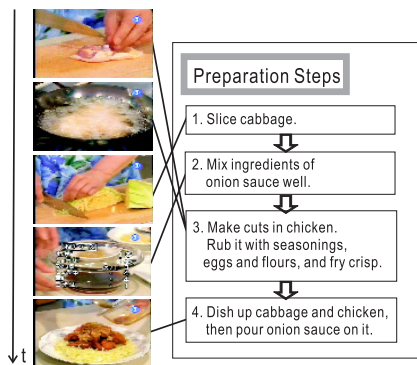


Fig. 1. Association of a cooking video and a recipe text.

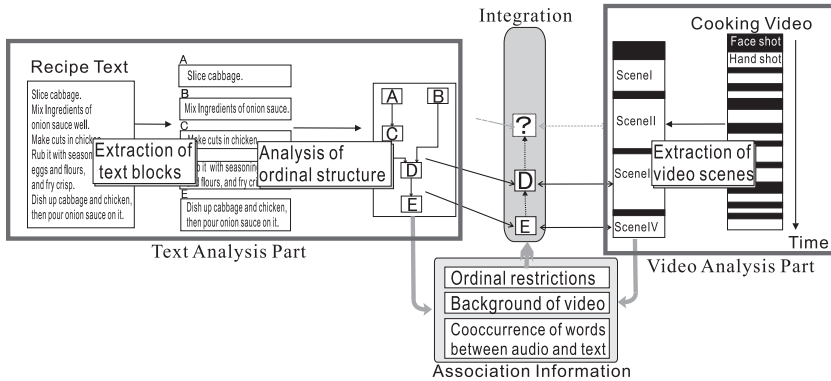


Fig. 2. Overview of the integration method.

Finally, association of each “text block” and “video scene” is performed. At this time, a text block with the highest relevance ratio to a video scene are matched together. The relevance ratio is calculated from the integration of information derived from multiple sources.

Each part of the system is explained in the following sections.

3 Media Analysis

3.1 Extraction of Text Blocks and Analysis of the Ordinal Structure

An overview of the text analysis¹ is shown in the left side of Fig. 2.

First, nouns, verbs and some modifiers are extracted from a cookbook, referring to a domain-specific dictionary. Especially, words which express a cooking condition (ex. “at high heat”) are important.

Next, a series of verbs starting with verbs which fulfill the following conditions is extracted as a “text block”.

1. Verbs which are associated to the same “ingredient noun”.
2. Verbs in a sentence which has “a container noun” + “in”.
(ex. “Bake onions in a frying pan.”)

At last, ordinal relations of text blocks are determined. We have already proposed a method to extract the verbal ordinal relation automatically, and have shown its effectiveness[1]. Using this method, ordinal restrictions of text blocks are extracted from the verbal ordinal relations.

¹ The entire procedure is for Japanese text.

3.2 Extraction of Video Scenes

“Video scenes” are extracted as groups of hand shots that have the same background.

First, cut detection is performed to a video sequence. In our implementation, we adopted a cut detection method using DCT clustering [2]. After the cut detection, the shots are classified into two categories; (1)Hand shot and (2)Face shot, as shown in Fig. 3. Hand and face shots are categorized automatically by face detection as described in our previous publication [3].

Next, hand shots are classified by background color distribution. Backgrounds are categorized into “board”, “table” and “range (=gas stove)” as shown in Fig. 3 (a) ~ (c), and “others” which are those that can not be categorized into any category.

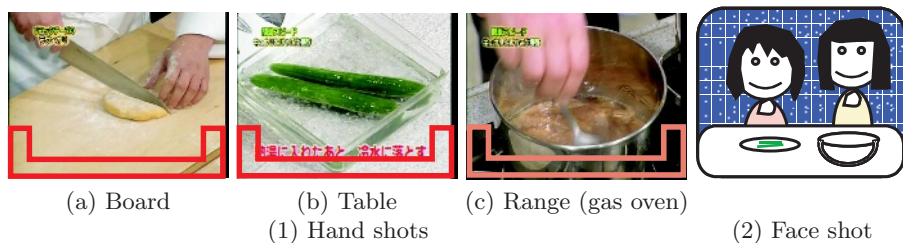


Fig. 3. Shot categories in cooking videos.

In this method, a supervised learning method using multiple cooking programs as training data is performed to extract which part of the images is most probable to represent the “background”. After that, hand shots are clustered using the color information of the “background” part. The “background” part which is extracted by this method is shown in Fig. 3 (a) ~ (c) as the emphasized block at the bottom. By this process, actual types of the background (ex. table or range) can not be specified, but shots with the same background could be distinguished.

Finally, a continuous series of hand shots with the same background is extracted as a “video scene”.

4 Multimedia Integration

In this section, the integration method which associates a “text block” and a “video scene” is explained.

As shown in Fig. 2, ordinal relations of “text blocks” is structured as an inverted tree. Therefore, when we analyze the restrictions in the order, it is easy to solve them from the latest one to the first one. A text block is associated with a “video scene” with the highest relevance ratio, from the last to the first.

Even for the same combination of a “video scene” and a “text block”, the “relevance ratio” between them may vary according to the order the association was tracked till then. Therefore a text block and a video scene are associated so that it maximizes the total score as a whole.

In this paper, a “relevance ratio” is defined as the total sum of the scores derived from the information listed below.

1. Ordinal restriction of text blocks.
2. Background information of the video scene.
3. The number of the words that cooccur in the text block and the audio data (closed caption = CC) of the video scene.

In order to adjust the influence of the above three information, scores X_1 , X_2 and X_3 which show the degrees of relation are assigned to each of them respectively. The score is distributed to each text block within the assigned score for each information.

The score calculation method to select a text block T_j associated with a video scene $S_{i=I}$ is explained below.

Information 1: Ordinal Restriction of Text Blocks

Here, a score is distributed so that text blocks, which may have more possibility to be associated with a scene $S_{i=I}$ according to the ordinal structure, should have higher score.

The following is an explanation of a score calculation method based on the example in Fig. 4. In Fig. 4, the latest two scenes (IV and III) have already been associated to text blocks, and the next scene (II) is waiting for the association.

The previously associated blocks (D and G) are defined as $T_{j \in \alpha}$. As shown in Fig. 4, we define the nearest blocks which are upper than $T_{j \in \alpha}$ blocks as the first candidates, and the nearest one which are upper than the first candidates as the second candidates. This is because the blocks lower than $T_{j \in \alpha}$ are scarcely possible to be associated to earlier scenes.

Let n_1 be the number of blocks of the first candidates. Then X_1/n_1 is the score of the text of the first candidates. The second candidate blocks get $X_1/n_1 \times n_2$, where n_2 is a suitable ratio ($1 \sim 2$).

Information 2: Background Information of the Video Scene

At first, background information of each scene can not be used as a hint for the association, because the type of the background is not specified yet. As the association proceeds, text blocks are associated to each video scene, the type of each background class could be inferred.

First, one of the four kinds of attributes in Tab. 1 are given to each word contained in a text block referring to a domain-specific dictionary.

The attribute of a text block is defined as the same as that of the words it contains. Although words with an attribute “o” are usually neglected, when a text block contains only words of attribute “o”, a text block will also have an attribute “o”.

Next, a process is selected among the following A \sim C according to the current state of the association process. (The background class of the target scene $S_{i=I}$ is defined as B_I .)

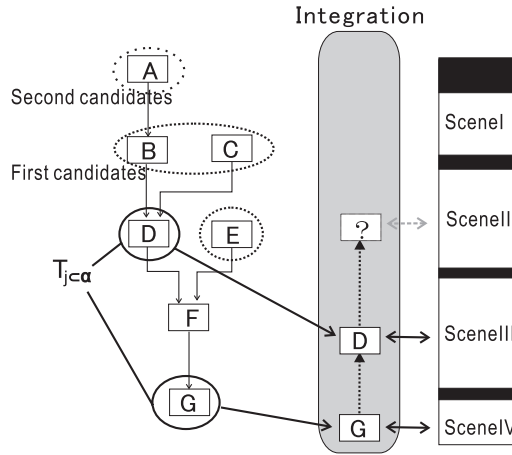


Fig. 4. Extraction of candidate text blocks using ordinal restrictions.

Table 1. Attributes of words related to video backgrounds.

attribute	property	associated background	examples
c	cut	board	cut, slice, knife, cutting board
h	heat	range	bake, heat, frying pan
m	work	table	dish up, mix, bowl, dish
o	others	—	add, chopsticks

A. There has been no scene associated with the text blocks: This is the first condition of the whole association process. A score 0 is given to all text blocks because no hint is available from the background information at this state.

B. Scene with background B_I has been associated with some text block: A score X_{ij} which shows the relevance ratio between a scene S_i and a text block T_j (attribute C_j) is given as in Eq. 1. The number of all text blocks which has been associated to scenes with a background class B_i is defined as n_i , and the number of text blocks with an attribute C_j among them is defined as n_{ij} .

$$X_{ij} = X_2 \times n_{ij} / n_i \tag{1}$$

C. Scenes with backgrounds except B_I is associated to a text block: X_{IJ} is the score between a text block $T_{j=J}$ (attribute $C_{j=J}$) and a video scene S_I . When the number of background types which has not been associated to a text yet (including B_I) is defined as n_B , X_{IJ} is defined as shown in Eq. 2. According to Eq. 2, for example, if many other background types have been associated with an attribute “h”, a scene with a background B_I will have a

lower relevance ratio with a text block with an attribute “h”.

$$X_{IJ} = \begin{cases} \frac{X_2 - \sum_{i \neq I} X_{iJ}}{n_B} & \text{if } \sum_{i \neq I} X_{iJ} < X_2 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Information 3: The Number of the Same Words in a Text Block and Audio Data (Closed Caption) of the Video Scene

High relevance ratio is distributed to a text block T_j when there are many common words between T_j and the closed caption contained in a video scene S_I . Here, the audio is introduced as a hint and the accuracy of the association is expected to improve.

First, effective words are extracted from all the closed captions in a video scene using a domain-specific dictionary. A relevance score X_{IJ} between a scene S_I and a text block T_j is defined as Eq. 3, when the number of common words between T_j and the closed caption in the scene S_i is defined as W_{ij} .

$$X_{IJ} = X_3 \times W_{IJ} / \sum_j W_{Ij} \quad (3)$$

5 Evaluation of the Integration

An evaluation experiment is performed according to the method described in the previous section. Experimental conditions are shown in Tab. 2. Parameters were determined manually through several trials.

The purpose of this experiment was to evaluate the integration method itself. Therefore, shot detection, categorization, and clustering using backgrounds were performed manually. On the text analysis part, extraction of text blocks and analysis of the ordinal structure were performed manually, too.

Video shots and text blocks were associated according to the method described in the previous section, and the result was compared with a ground truth which was created manually.

Table 2. Experimental conditions.

(a) Features of data			(b) Parameters	
cooking program	the number of recipes	duration	parameter	value
“K”	10	1’24”	X_1	60
“O”	10	1’18”	X_2	60
Total	20	2’42”	X_3	100

Table 3. The result of evaluation experiment (%).

Used Informations	“K”	“O”	Average
1. Ordinal restriction	20.2	20.5	20.4
2. Background info.	24.6	20.5	22.6
3. Closed caption	60.5	58.9	59.7
All (Proposed method)	83.3	74.1	78.8

The result is shown in Tab. 3. In this table, the results using the Informations 1, 2 and 3 individually are shown to be compared with the result of the proposed method to evaluate the effectiveness of the integration of multimedia information proposed in this paper.

The result shows that the accuracy by the proposed method is much higher than the accuracy using only one information source, and the total average accuracy is about 80%. Through this, the effectiveness of the proposed method is shown.

6 Conclusion

We have been working on integration of video with supplementary documents, such as cooking programs. We propose an integration system that performs semantic segmentations of video and text, and associate them together. This association was realized using the ordinal restriction in a recipe, cooccurrence of words in the text and the audio in the video, and the relation between the background in a video and words which describe the situation in the text.

We introduced the result of an evaluation experiment and showed the effectiveness of the proposed integration method. Through our method, many applications should become possible, such as a cooking navigation software.

Acknowledgement. Part of the work presented in this paper was supported by the Grant-in-Aid for Scientific Researches (14380173) from the Japanese Society for the Promotion of Science.

References

1. Reiko Hamada, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka: "Structural Analysis of Cooking Preparation Steps in Japanese", Proc. Fifth Intl. Workshop on Information Retrieval with Asian Languages *IRAL2000*, pp.157-164 (Oct. 2000)
2. Y. Ariki, Y. Saito: "Extraction of TV News Articles Based on Scene Cut Detection", *Proc. ICIP'96*, pp.456-460 (1996)
3. Koichi Miura, Reiko Hamada, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka: "Motion Based Automatic Abstraction of Cooking Videos", Proc. ACM Multimedia 2002 Workshop on Multimedia Information Retrieval, (Dec. 2002)
4. H. D. Wactlar, A. G. Hauptmann, M. G. Christel, R. A. Houghton, A. M. Olligschlaeger: "Complementary Video and Audio Analysis for Broadcast News Archives", *Comm. ACM*, Vol.45, No.2, pp.42-47 (Feb. 2000)

Teleconference System with a Shared Working Space and Face Mouse Interaction

Jin Hak Kim, Sang Chul Ahn, and Hyoung-Gon Kim

Imaging Media Research Center, KIST
39-1 Hawolgok-dong, Sungbuk-gu, Seoul KOREA 136-791
{kjh, asc, hqk}@imrc.kist.re.kr
<http://www.imrc.kist.re.kr>

Abstract. This paper presents a new computer-based teleconference system. In order to build the system, we modified VNC(Virtual Network Computing) program and implemented *Face mouse* system. *Face mouse* is a mouse pointer that is followed by face video. It has multiple functions of video communication channel, mouse identification, and a connection between face window and a shared working space. By moving *Face mice* across a shared working space, the proposed system can provide effective teleconference environment. The proposed system has advantages of providing a larger shared working space and more intuitive interaction than current teleconference systems.

1 Introduction

Teleconference enables people to have a meeting together on a subject from remote locations. Teleconference can increase productivity by removing waste of travel time and delay of decision-making, which could occur when insisting on a face-to-face meeting. However, a teleconference system has various requirements to provide natural and easy-to-use teleconference environment. First of all, a teleconference system should provide various communication channels to increase naturalness of conference. Voice and video channels are examples of them. Secondly, a teleconference system should provide a shared working space as well. In a face-to-face meeting, people usually sit around a table, and discuss over or work on some materials on it. The table plays a role of a shared working space where people can share their data and work collaboratively. Teleconference also requires such a space. Thirdly, a teleconference system should provide effective interaction methods in the shared working space. In a face-to-face meeting, people use their hands for interaction on a table. They can easily indicate an item on a table and hand over a material by hands. People in a teleconference need to do similar things for effective meeting.

As computer technology advances, computer-based teleconference system has been studied by a lot of researchers. Most of studies have focused on efficient provision of video and audio communication channels. [1], [2], [3] are some examples of such study. However, most of the teleconference systems provide only

whiteboard or blackboard style shared working spaces. They don't provide intuitive interaction method with a shared working space. Situation is the same even for commercial teleconference systems. NetMeeting [4] of Microsoft, that is widely used, shows the typical configuration of teleconference system of these days. It has a video window for visual communication, and provides a whiteboard program for collaboration. With NetMeeting, teleconference is far from face-to-face meeting, and possible interactions are limited. This paper proposes a new computer-based teleconference system that provides people with a computer screen as a shared working space and more intuitive interaction with it.

2 System Overview

The proposed system uses a computer screen as a shared working space for teleconference. Fig. 1 shows the configuration of the proposed system. It shows that clients share the server screen at the same time. Here we call 'Server' a computer that provides its screen as a shared working space, and 'Client' a computer that shows remote server screen to user. Not only the user of server but also users of client computers can control a server computer by their mice.

In order to provide such a shared working space, we need remote computer sharing. There are several programs for remote computer sharing such as VNC, pcAnywhere, Windows Terminal Service. But, we decided to use and modify VNC (Virtual Network Computing) program because VNC is open source and has been stabilized for a long time. VNC was originally developed by Olivetti Research Laboratory (ORL) in Cambridge, UK [5], [6]. VNC uses Remote Frame Buffer (RFB) Protocol [7], and is divided into a server part and a client part. Using VNC, several users can see the same screen and data that they want to share. Therefore, VNC is appropriate for providing a shared working space.

Since VNC offers multi sessions at the same time, several users can connect to a server and control the mouse pointer of the server. In this case, however, VNC provides only sharing of a mouse for multiple users. So, if several users operate the mouse of a server through VNC, it can be very confusing because users cannot know who is controlling the mouse. We need to provide multiple mouse pointers and a discrimination method for effective teleconference and collaboration. Thus, we modified VNC and implemented so-called '*Face mouse*' to solve the identification problem of mouse pointers along with the problem of video transmission. *Face mouse* is a mouse followed by a video that is showing a face. Fig. 1 shows that faces of users are displayed on the screen of the server computer. They are the *Face mice* that act as mice while they are showing user's faces. As can be seen in Fig. 1, each client sends face video of its user to the server, and the server composites the face video on its screen. The face video is used for the purpose of mouse identification as well as video communication at the same time. Since *Face mouse* is a mouse, its position is dynamically controlled by user. It is different from existing teleconference systems where face videos are located at a fixed position. By moving face videos, the proposed system has advantage of using whole computer screen as a shared working space rather than using a

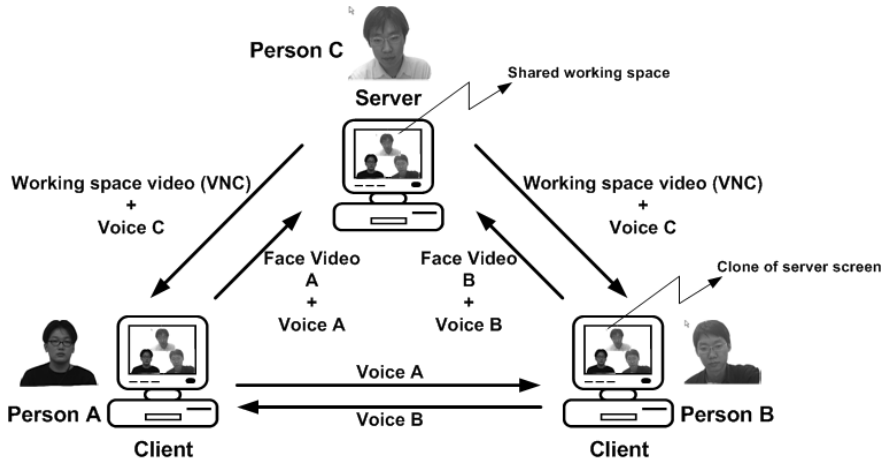


Fig. 1. Configuration of the proposed teleconference system

part of it. After the server processes all, VNC enables remote users to share and see the screen of server computer that all the *Face mice* are composited on. For the voice communication channel, the proposed system uses ‘Speak freely’ that is another open source program.

3 Face Mouse

Face mouse has actually 3 roles: video communication, mouse identification, and more intuitive interaction. Video communication and mouse identification are accomplished by face videos as mentioned before. Additionally, the proposed system tries to connect face videos and a shared working space. In a face-to-face meeting, users and a shared working space(eg. meeting table) are not separated. Users can access a table and do any interaction on the table such as picking up a material, moving it, and pointing an item by hand. On the contrary in the existing teleconference systems, display region of face videos and a shared working space, if exists, are separated. Fig. 2 shows screen shots of existing teleconference systems.

Users in the videos cannot access directly a shared working space. Though users are provided with mouse pointers in the existing teleconference systems, there is no correlation cue to connect face and user’s mouse pointer. When a user is pointing an item, other users cannot identify intuitively who is pointing. *Face mouse* tries to alleviate the situation by attaching face video to a mouse and by moving face video over a shared working space. In the proposed system, a user can use even gesture to point an item after locating his *Face mouse* around it. This resembles real situation in a face-to-face meeting. Thus, *Face mouse* enables more intuitive interaction in teleconference and may draw more attention from other people.

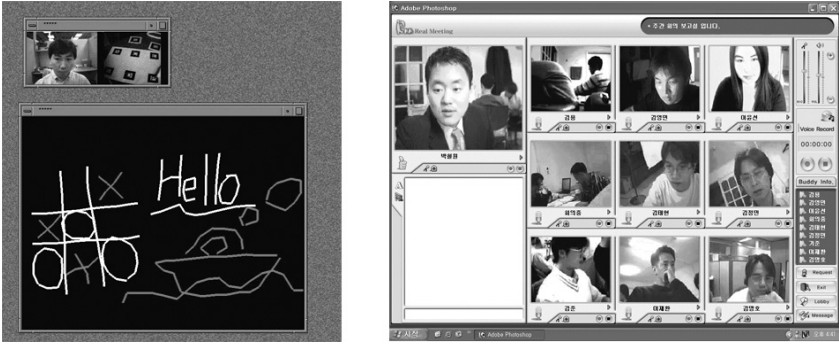


Fig. 2. Screen shots of existing teleconference systems

3.1 Implementation Overview

Face mouse can be implemented only with a small web camera and networked computers. *Face mouse* was implemented using C++ on Windows platform. Fig. 3 shows a simple flowchart of *Face mouse* program. Client computer captures live face video with a camera and applies background subtraction operation to get a foreground face video. Though it is not shown in the flowchart, background information is gathered for initial 30 frames. The background subtraction is not a mandatory operation but an optional operation for quality face video. Then, it delivers to a server computer foreground face video along with mouse position of the client computer. When the server computer receives the video, it makes background area transparent and leaves foreground face. Then, it locates the face video according to the received mouse position. If the mouse is not moved, the location of face video is maintained as before.

3.2 Client Side Implementation

The proposed system uses foreground face video for *Face mouse*. Background subtraction is not mandatory, but advantageous because it reduces data size and makes more attention go on a foreground face. It also enables more working space to be seen rather than useless background video after composited on the screen. We adopted simple background subtraction method to reduce computational overhead. For the background subtraction, background information is gathered for initial 30 frames. First, we normalize R , G , B data and get r , g , b for each pixel to reduce the variation due to light change[8]. Eq. (1) shows the normalization process for R component, for instance.

$$r(x, y) = \frac{R(x, y)}{R(x, y) + G(x, y) + B(x, y)}, \text{ for } 0 \leq x < M, 0 \leq y < N, \quad (1)$$

where M and N represent width and height of video, respectively. Then, it computes averages r_m, g_m, b_m and standard deviations r_s, g_s, b_s of normalized r ,

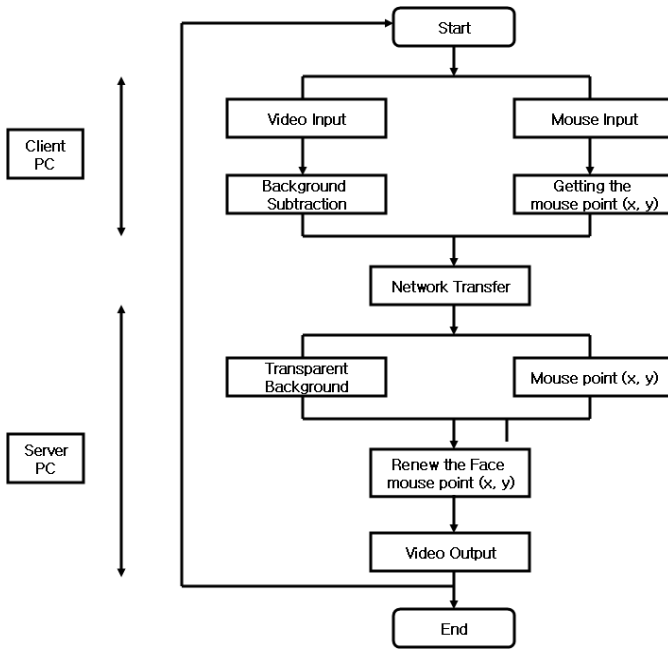


Fig. 3. Flowchart of Face mouse system

g , b data for 30 frames. Once they are obtained, these averages and standard deviations are used to distinguish background pixel from foreground pixel as follows.

$$f(x, y) = \left\{ \begin{array}{l} Bg, \text{ if } r_d(x, y) \leq kr_s(x, y) \text{ and } g_d(x, y) \leq kg_s(x, y) \\ \quad \text{and } b_d(x, y) \leq kb_s(x, y) \\ Fg, \text{ otherwise} \end{array} \right\}, \quad (2)$$

where k is a predetermined constant and r_d, g_d, b_d are differences between averages and new pixel values, respectively. For instance, r_d is given by $r_d(x, y) = |r_m(x, y) - r(x, y)|$. After background subtraction, foreground face video is sent to a server computer along with mouse position data.

3.3 Server Side Implementation

A server computer provides its screen as a shared working space. On the shared working space, the server computer overlays face videos received from client computers. Before displaying them, it makes background areas transparent in face videos. Additionally, it adds mouse icon to the top-left side of each face video so that it looks like a mouse pointer. Still it is not an actual mouse since it does not have control of computer. The control of computer can be provided by VNC. VNC enables remote users to share mouse control of a local computer.

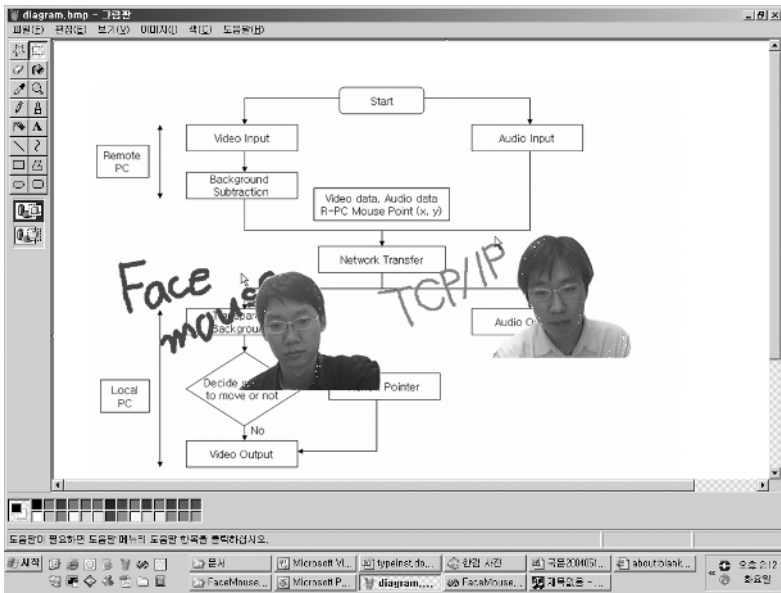


Fig. 4. A screen shot of Face mouse

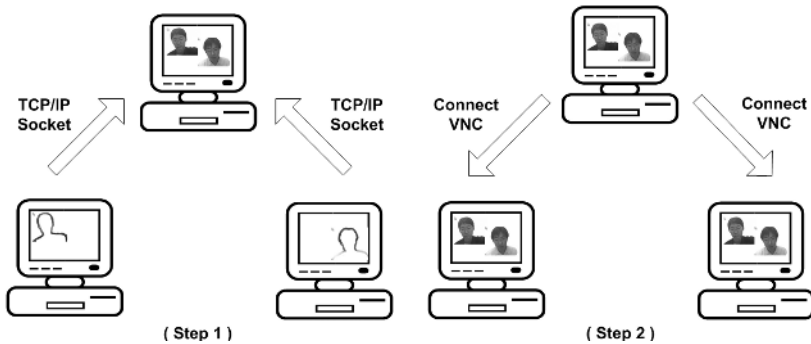


Fig. 5. Face mouse composition and visual feedback by VNC

However, when several users try to control the mouse at the same time, they compete each other, and motion of the mouse may become confusing. Therefore, we removed mouse motion control function from VNC. Then, we modified VNC so that it only receives click events from *Face mouse* program. When modified VNC receives a click event, it moves the system mouse to a relevant position and does a click action.

Now, the client program, the server program, and the modified VNC function cooperate each other and form the *Face mouse* program. Fig. 4 shows a screen shot of server computer where *Face mouse* is working. After *Face mouse* is

overlaid on the server screen, the screen is transferred to client computers by VNC. So, users of client computers can receive visual feedback of their control on a shared working space. Fig. 5 shows the process.

4 Conclusion

Teleconference has potentially wide application area. Recently some of teleconference systems have started to be used, but there exist limitations in their functions. Teleconference is still far from face-to-face meeting with current systems. This paper proposed a new computer-based teleconference system that can provide people with a computer screen as a shared working space and more intuitive interaction with it. In order to build the system, we implemented *Face mouse* system. *Face mouse* system provides multiple mouse pointers that are followed by face videos. *Face mouse* supports the roles of video communication as well as mouse identification. It also connects naturally face windows with a shared working space by attaching face videos to mouse pointers and by moving it across a shared working space. The proposed system has advantages in the points that people can use larger shared working space and more intuitive interaction for teleconference.

References

1. Alan Piszcz, Eddie Cheung, David Debarr, Nicholas Orleans: Realizing a Desktop Collaborative Workspace. The MITRE Corporation, McLean, VA 22102. (1998)
2. Mark Billinghurst, Hirokazu Kato: Collaborative Mixed Reality. Proc. of First International Symposium on Mixed Reality (ISMR '99). Springer-Verlag Berlin (1999) 261-284
3. Ahmet Uyar, Wenjun Wu, Hasan Bulut, Geoffrey Fox: An Integrated Videoconferencing System for Heterogeneous Multimedia Collaboration. Department of Electrical Engineering and Computer Science, Syracuse University.
4. NetMeeting. Microsoft Corporation.
<http://www.microsoft.com/windows/netmeeting>
5. Stafford-Fraser, Q., Weatherall J.: Virtual Network Computing. AT&T Laboratories Cambridge, <http://www.uk.research.att.com/archive/vnc/>
6. Richardson, T., Stafford-Fraser, Q., Wood, K., Hopper, A.: Virtual Network Computing, IEEE Internet Computing, Vol. 2, No. 1, (1998)
7. Tristan Richardson, Kenneth R. Wood. The RFB Protocol. ORL Cambridge. Version 3.3 (1998)
8. Hyoung-gon Kim, Nam Ho Kim, and Sang Chul Ahn: Skin Region Extraction using Moving Color Technique. Proc. of ISPACS'98 (1998) 73-77

Performance Analysis for Serially Concatenated FEC in IEEE802.16a over Wireless Channels

Kao-Lung Huang and Hsueh-Ming Hang

Department of Electronics Engineering, National Chiao Tung University,
Hsinchu 300, Taiwan, R.O.C.,
{u8911839, hmhang}@cc.nctu.edu.tw

Abstract. To design an IEEE802.16a wireless transmission system, we investigate the performance of its serially concatenated forward error correction (FEC) system for both additive white Gaussian noise (AWGN) channel and the fully interleaved Rayleigh fading channel (FI-RFC). We derive the union upper bounds on the bit error rate (BER) for the Reed-Solomon (RS) code and the rate-compatible punctured convolutional codes (RCPC). These theoretical bounds are compared with the simulation results and it shows that the derived bounds are fairly tight. In addition, simulation results of the packet error rate (PER) are also presented. We thus in this paper provide researchers a suitable operational range of IEEE802.16a in terms of signal-to-noise ratio and acceptable performance.

1 Introduction

Recently, IEEE has proposed the standard referred to as IEEE802.16a for the local and metropolitan area network [1]. Its serially concatenated FEC scheme consists of an RS(255,239,8) code as the outer code and RCPC codes as the inner code.

To combat the severe channel degradation, concatenating RS code with convolutional code (CC) could enhance their error control performance [2]. One advantage of using RS/RCPC concatenated codes is that they can provide multiple services and multiple rate transmissions, which is particularly useful for multimedia communications.

The idea of RCPC codes was first introduced by Hagenauer [3]. The performance analysis of this type of codes over wireless channels could be found in [4]. However, few studies have been reported on the performance analysis of the concatenation of the RS code and the RCPC codes together. In addition, we like to identify a suitable operational range in terms of signal-to-noise ratio and acceptable performance. The aim of this paper is to investigate the performance of RS/RCPC concatenation defined by the IEEE802.16a specifications over the AWGN and the FI-RFC channels [4]. We derive the union upper bounds on the BPSK-modulated BER (bit-error-rate) at the output of the concatenated RCPC and RS code. Also, we compare the theoretical bounds with the simulation results.

The rest of this paper is organized as follows. Sec. 2 describes the system model. Union upper bounds on BER and PER are derived in Sec. 3. Sec. 4 shows the simulation results and comparisons are made with the theoretical upper bound. Finally, conclusions are drawn in Sec. 5.

2 System Model

The model of the transmission system to be analyzed is shown in Fig. 1. The message bit stream in the analysis and simulation is assumed to be a random bit sequence u_i . The message bits are packed into blocks of 239x8 bits since the RS code operates over $GF(2^8)$. Each block is first coded by RS(255,239,8) coder. This coder inserts 16x8-bit redundancies for each block. Thus, the output is a packet of length 255 bytes, which are then fed into the RCPC coder. The mother code of this RCPC code has a coding rate of 1/2 and a constraint length of 7. With different puncture patterns (perforation matrices), this RCPC is capable of producing five different coding rates: 1/2, 2/3, 3/4, 5/6, and 7/8. Therefore, depending on the channel condition, we can select an appropriate bit rate that leads to the best trade-off between data throughput and error probability. A number of tailing bits are inserted to ensure proper decoding operation with an acceptable decoding delay. The choice of tailing bits leads to a tradeoff between the error control performance and the data throughput rate. Another parameter in the practical Viterbi decoding process is the decoder depth (decision path length). This study uses 10 tailing bits and the decoder depth is 210 bits, 35 times the memory span of the mother convolutional code of the given RCPC to get better performance in the Rayleigh fading channel. The coded bits are the BPSK-modulated and the modulated symbols $\{x_i\}$ are transmitted.

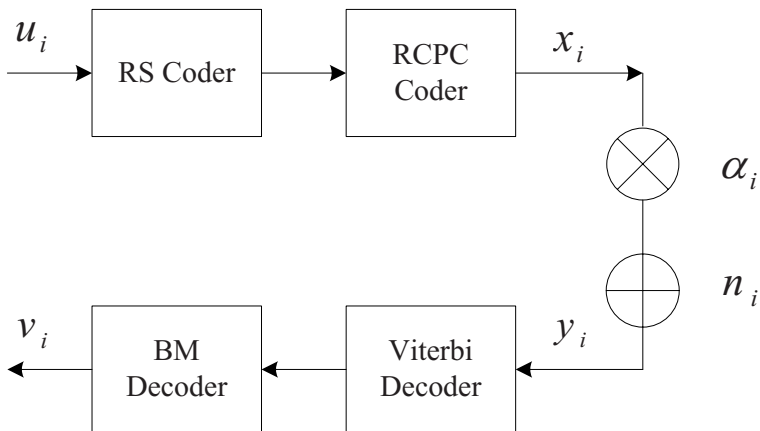


Fig. 1. System Model.

Suppose the test channel has slow and flat Rayleigh fading, then the phase error can be perfectly tracked. We further assume coherent demodulation is available. The received sample y_i is thus expressed in the form

$$y_i = \sqrt{\frac{2R_c E_b}{N_0}} \alpha_i x_i + n_i \quad (1)$$

where E_b is the energy per information bit, R_c is the selected code rate of RCPC codes, and n_i is the zero mean white Gaussian noise sample with unit variance. Because we assume the channel is a fully interleaved Rayleigh fading channel, the sequence $\{\alpha_i\}$ of the fading envelope is independent and is identically Rayleigh distributed with the following probability density function,

$$f(\alpha) = \frac{\alpha}{\sigma^2} e^{-\frac{\alpha^2}{2\sigma^2}} \quad (2)$$

where σ^2 is the time-average power of the received signal before the envelop detection and the fading envelope α has the properties of $\alpha \geq 0$ and $E[\alpha^2] = 1$. Moreover, α is set to one for the AWGN channel. Soft decision decoding (SDD) with no quantization is used in conjunction with the Viterbi decoding process. In the RS decoding, the Berlekamp-Massey based algorithm is used in simulation and only the error-correction ability is considered. After RS decoding, the output bit sequence $\{v_i\}$ is obtained and then the BER (bit error rate) and PER (packet error rate) of the RS code are calculated.

3 Performance of Serially Concatenated FEC

In this section, the upper bounds on BER and PER of serially concatenated FEC are derived for the AWGN and the FI-RFC channels. In the IEEE802.16a specifications, the concatenated FEC can operate together with both BPSK and QPSK modulations; however, we analyze the BPSK modulated signal only, since the performance of QPSK modulation is essentially the same as that of BPSK [6].

3.1 Union Upper Bound on BER of RCPC Codes

The typical union bound can be expressed in the form

$$P_{b,RCPC} \leq \frac{1}{pp} \sum_{d=d_f}^{\infty} c_d p_2(d) \quad (3)$$

where pp , d_f , and c_d are the puncture period, free distance, and weight distribution coefficient, respectively. Detailed description of the above parameters can be found in [5]. Here, we calculate the upper bound of $P_{b,RCPC}$ by summing up the Hamming distance d in the range of d_f to $d_f + 9$. Without loss of generality, transmission of the all-zero sequence is assumed and the pair-wise error probability $p_2(d)$ is referred to the case of selecting an incorrect path with Hamming

weight d in the Viterbi decoding process. In AWGN and coherent BPSK scheme, $p_2(d)$ is given by

$$p_2(d) = Q\left(\sqrt{\frac{2dR_cE_b}{N_0}}\right) \quad (4)$$

where the Q is defined by $Q(x) = (\sqrt{2\pi})^{-1} \int_x^\infty e^{-\frac{x^2}{2}} dx$. Under the FI-RFC, coherent BPSK, SDD, and perfect channel estimates $\{\alpha_i\}$ assumptions, the pair-wise error rate $p_2(d)$ is derived using the concept of diversity. The concatenated FEC utilizes the time diversity technique to attain independent fading envelope α_i among the received coded symbols. A closed form of the pair-wise error rate $p_2(d)$ is given by

$$P_2(d) = \gamma_s^d \sum_{i=0}^{d-1} \binom{d-1+i}{i} (1-\gamma_s)^i \quad (5)$$

where $\gamma_s = 0.5\left(1 - \sqrt{\frac{R_c\gamma_b}{1+R_c\gamma_b}}\right)$ and $\gamma_b = \frac{E_b}{N_0}$ [6]. This probability is the d^{th} order order time diversity, which is equivalent to the d^{th} order path diversity when the max ratio combining is applied.

3.2 Union Upper Bound on BER of RS Codes

The union upper bound on PER can be constructed based on the symbol error rate, which is derived from the union bound on the BER of RCPC. A union upper bound on the PER of the RS code is expressed in the form

$$P_{p,RS} \leq \sum_{i=t_c+1}^n \binom{n}{i} p^i (1-p)^{n-i} \quad (6)$$

where $p = 1 - (1 - P_{b,RCPC})^8$. When a RS code word error occurs, the associated 8-bit symbol error rate is given by

$$P_{s,RS} \leq \frac{1}{n} \sum_{i=t_c+1}^n i \binom{n}{i} p^i (1-p)^{n-i} \quad (7)$$

and the corresponding upper bound on BER is approximated by

$$P_{b,RS} \leq \frac{2^{k-1}}{2^k - 1} P_{s,RS} \quad (8)$$

4 Simulation Results

In this section, the average BERs of RCPC and RS codes are calculated using the simulated data. The BER values are plotted and compared to the corresponding theoretical upper bounds for the code rates 1/2, 2/3, 3/4, 5/6, and 7/8. In the simulation, up to 10^7 data bits are transmitted. Simulations are plotted on Fig. 2, Fig. 3, Fig. 4, and Fig. 5, where T and S denote theoretical bound

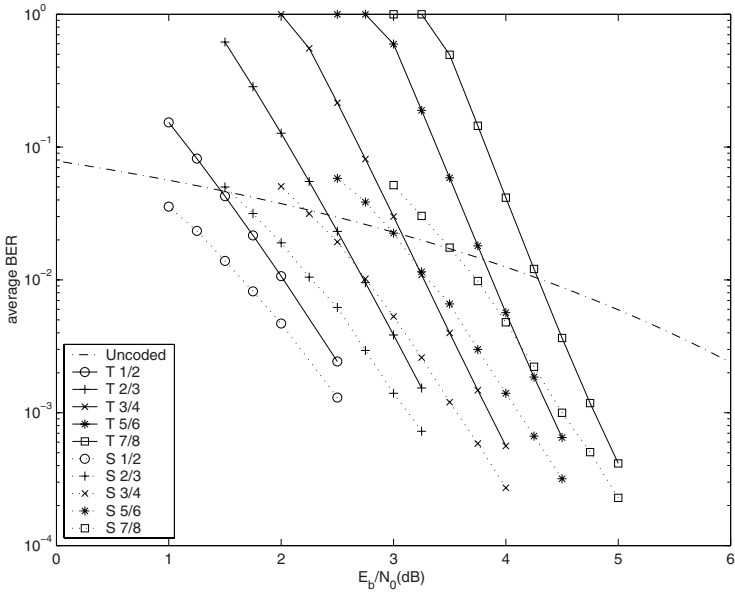


Fig. 2. BER of RCPC codes in AWGN.

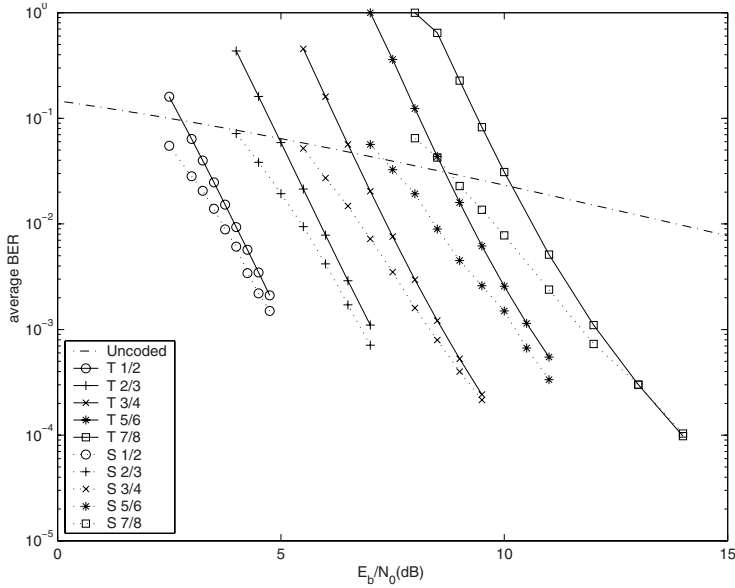


Fig. 3. BER of RCPC codes in FI-RFC.

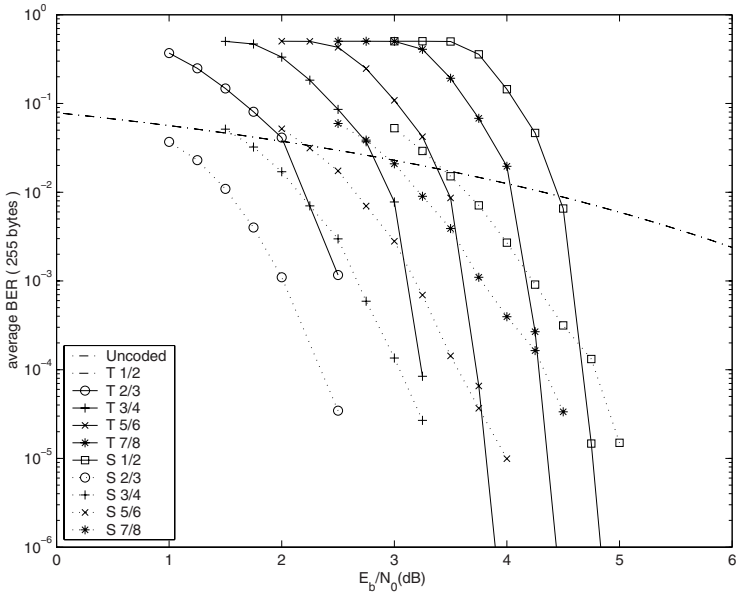


Fig. 4. BER of RS codes in AWGN.

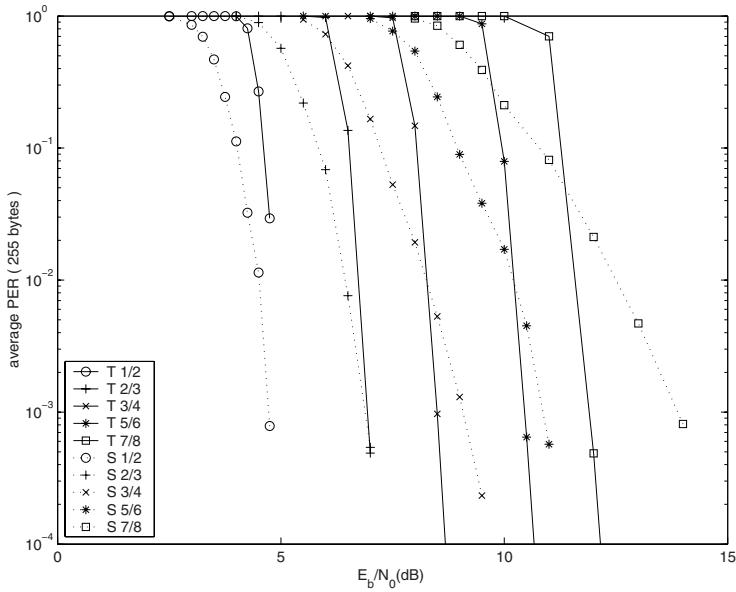


Fig. 5. BER of RS codes in FI-RFC.

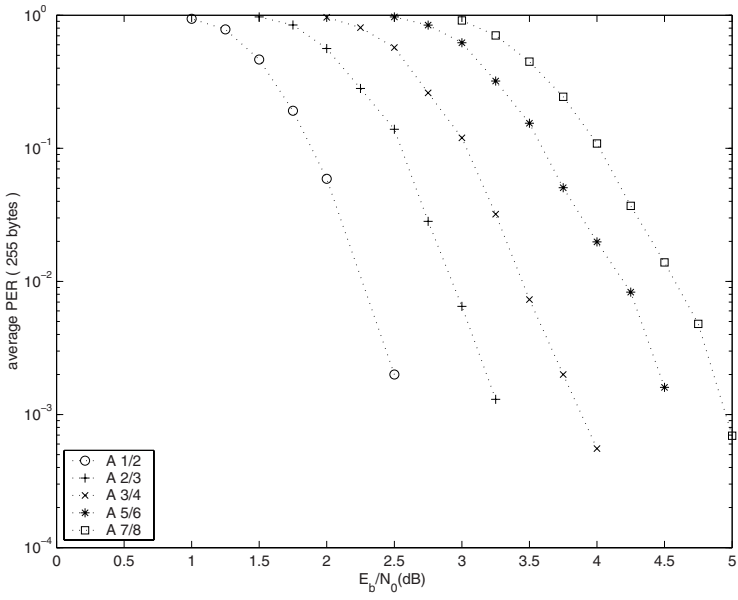


Fig. 6. PER of RS code in AWGN.

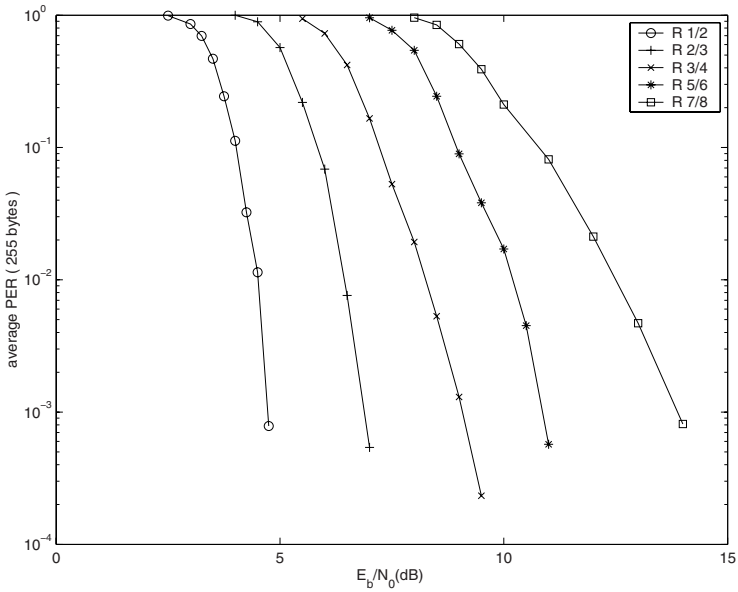


Fig. 7. PER of RS code in FI-RFC.

and simulation results respectively. Simulation results of BER are close to the theoretical bounds, particularly when the signal-to-noise ratio is large. These results are consistent with the findings of the previous study [6]. Also, it shows that simulation results and theoretical upper bounds are fairly tight. In addition, the uncoded BER curve is also plotted. It is clearly shown that the coding gain can be obtained if the signal-to-noise ratio is sufficiently large.

Fig. 6 and Fig. 7 showed the average PER of 255-byte RS packets in the AWGN and FI-RFC channels, where A and R denote AWGN and FI-RFC channels respectively. With this data, researchers who are primarily interested in packet-based transmission might simulate their testing platform more easily with IEEE802.16a specification and get more realistic results.

5 Conclusions

In this paper, the performance of the serially concatenated CFEC defined by the IEEE802.16a specifications is analyzed and simulated for both AWGN and RFC channels. The RFC channel is assumed to be slow and flat fading and is fully interleaved. Moreover, the soft decision Viterbi decoding has no quantization. The upper bounds on BER of RCPC and RS codes have been derived and are compared to the simulation results. We thus found that the upper bounds are quite tight.

Also, the PER performance is simulated and summarized. With this set of data, researchers interested in packet-based transmission could easily design their IEEE802.16a systems that meet the target performance. In conclusion, we provide a suitable operational range for IEEE802.16a, which is a trade-off between the signal to noise ratio and the desired performance.

References

1. The Institute of Electrical and Electronics Engineers, "IEEE standard for Local and metropolitan area networks," Apr. 2003.
2. G. Solomon and H. C. A. van Tilborg, "A connection between block and Convolutional Codes," *Slam J. Appl. Mathematics*, vol. 37, no. 2, pp. 358–369, Oct. 1979.
3. J.Hagenauer, "Rate-Compatible Punctured Convolutional Codes(RCPC Codes) and their applications," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 389–400, Apr. 1988.
4. J.Hagenauer, "Viterbi Decoding of Convolutional Codes for Fading- and Burst-Channels," *Proc. Of the Zurich Seminar on Digital Communications*, 1980.
5. G. Begin, D.Haccoun, and C. Paquin, "Further Results on High-Rate Punctured Convolutional Codes for Viterbi and Aequential Decoding," *IEEE Trans. Commun.*, vol. 38, no. 11, pp. 1922–1928, Nov. 1990.
6. J. G. Proakis, *Digital Communication, 3rd ed.* New York: McGraw-Hill, 1995.

Successive Interference Cancellation for CDMA Wireless Multimedia Services

Jin Young Kim¹ and Yong Kim²

Department of Radio Science and Engineering
Kwangwoon University, Seoul, Korea
jinyoung@daisy.kw.ac.kr, kracon@hanmail.net

Abstract. In this paper, successive interference cancellation (SIC) scheme is proposed for CDMA wireless multimedia services. The bit error probability of the proposed system is simulated over both AWGN and Rayleigh fading channels. For a multimedia transmission model, a multicarrier transmission strategy and a variable spreading length (VSL) scheme are proposed. The comparative performance with different data rates is investigated and simulated. The results of the paper can be applied to the design of the fourth-generation mobile communication system with wireless multimedia traffics.

1 Introduction

In a conventional DS-CDMA system, a particular user's signal is detected by correlating the entire received signal with that user's code waveform[1]. A conventional DS-CDMA system treats each user separately as a signal while the other users are considered as interference, i.e., multiple access interference (MAI). The existence of MAI has a significant impact on the performance and capacity of conventional DS-CDMA systems. To improve the detection performance of each individual user, multiuser detection has been proposed [2, 3]. Multiuser detection also referred to as joint detection or interference cancellation. Code and timing information of multiple users is jointly used to better detect each individual user.

In this paper, successive interference cancellation (SIC) detector scheme is proposed for a multirate multicarrier DS-CDMA system and its performance is analyzed and simulated. The bit error probability of the SIC detector is simulated and the comparative performance with different data rates is investigated.

The paper is organized as follows: In Section 2, the multirate multicarrier DS-CDMA system is modeled. In Section 3, the transmitter, channel, and receiver models are described. In Section 4, the SIC cancellation algorithm is described. Simulation results are presented and discussed in Section 5. In Section 6, conclusions are drawn.

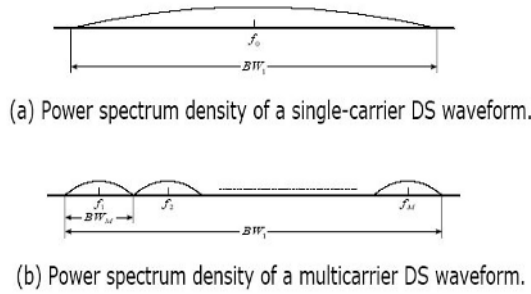


Fig. 1. Power spectrum density of DS waveform.

2 Multirate Multicarrier DS-CDMA System

2.1 Multicarrier DS-CDMA System

Fig. 1(a) shows a bandlimited single-carrier wideband DS waveform in the frequency domain, where the bandwidth, BW_1 , is given by

$$BW_1 = (1 + \alpha) \frac{1}{T_c}. \tag{1}$$

In (1), $0 \leq \alpha \leq 1$, and T_c is the chip duration of the single-carrier system. In a multicarrier system, we divide BW_1 into M equi-width frequency bands as shown in Fig. 1(b), where all bands are disjoint. Then the bandwidth of each frequency band, BW_M , is given by

$$BW_M = \frac{BW_1}{M} = (1 + \alpha) \frac{1}{MT_c}. \tag{2}$$

Note that MT_c is the chip duration of the multicarrier system, and M is the number of carriers. In a multicarrier system, carrier frequencies are usually chosen to be orthogonal to each other, i.e., carrier frequencies satisfy the following condition:

$$\int_0^{T_c} \cos(\omega_i t + \phi_i) \cos(\omega_j t + \phi_j) dt = 0, i \neq j, \tag{3}$$

where T_c is the chip duration, ω_i and ω_j are, respectively, the i th and j th carrier frequencies, and ϕ_i and ϕ_j are arbitrary carrier phases, respectively. This is done so that a signal in the j th frequency band does not cause interference in the correlation receiver for the i th frequency band.

2.2 Multirate System

The future mobile communication systems are expected to support several kinds of communication services, including, e.g., voice, image, and video transmission.

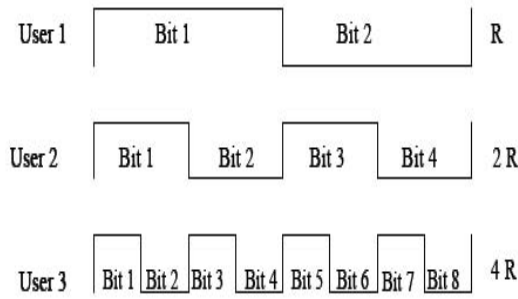


Fig. 2. Multirate strategy.

Multi-code (MC) access and variable-spreading-length (VSL) access are two widely applied realizations for multirate services in DS-CDMA system. In VSL systems, transmitting the data of user who has higher rate is realized by assigning shorter spreading codes, which also keeps the bandwidth fixed. The spreading factors in VSL systems are various from user to user by their data rates. The higher data rate is multiple of the basic (lowest) data rate. In Fig. 2, the VSL strategy is shown for multimedia traffic.

3 System Model

3.1 Transmitter Model

The transmitter for the k th user is shown in Fig. 3. The data sequence $b_k(t)$ is multiplied by spreading sequence $c_k(t)$ of period N , and modulated by the M subcarriers. The transmitted signal of the k th user is given by

$$s_k(t) = \sqrt{2P}b_k(t)c_k(t) \sum_{m=1}^M \cos(\omega_m t + \theta_{k,m}), \tag{4}$$

where P is transmit power per subcarrier, ω_m is the m th subcarrier frequency, and $\theta_{k,m}$ is the m th subcarrier phase of k th user. The transmit power per subcarrier is given by $P = \frac{E_b}{MT_b}$ where E_b is energy per bit and T_b is bit duration.

3.2 Channel Model

The frequency-selective multipath fading channel model is shown in Fig. 4. From the tapped delay line model for a frequency-selective fading model, the channel impulse response for the user k is given by

$$h_k(t) = \sum_{l=0}^{L-1} f_{k,l}(t)\delta_k(t - lD), \tag{5}$$

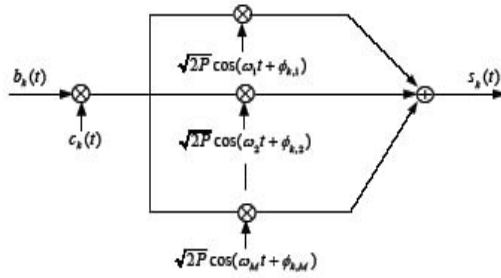


Fig. 3. Block diagram of transmitter for the k th user.

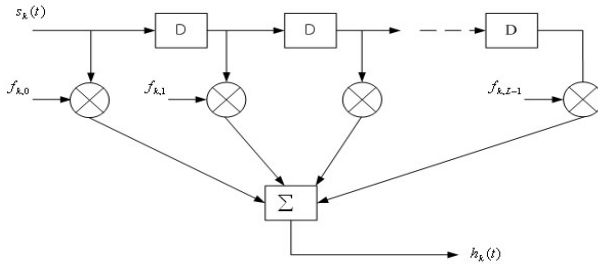


Fig. 4. Frequency-selective multipath fading channel model.

where L is the number of multipaths, D is tap spacing, and $f_{k,l}$ is the path gain and has a Rayleigh distribution. The number of multipaths for each user is given by

$$L = \lfloor \frac{D_s}{D} \rfloor + 1, \tag{6}$$

where D_s is multipath delay spread and $\lfloor x \rfloor$ is the largest integer contained in x . Since a data bit duration is much larger than channel delay spread in this multirate multicarrier DS-CDMA system, intersymbol interference (ISI) can be neglected.

3.3 Receiver Model

For asynchronous system with K users, the received signal is given by

$$r(t) = \sqrt{2P} \sum_{k=1}^K b_k(t - \tau_k) c_k(t - \tau_k) \sum_{m=1}^M \alpha_{k,m} \cos(\omega_m t + \phi_{k,m}) + n(t), \tag{7}$$

where τ_k is propagation delay uniformly distributed in $[0, T_b)$, $\alpha_{k,m}$ is Rayleigh-distributed fade envelope, $\phi_{k,m}$ is phase for the m th subcarrier of the k th user,

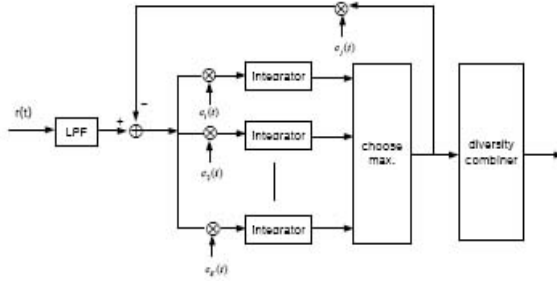


Fig. 5. Block diagram of receiver for interference cancellation scheme.

and $n(t)$ is additive white Gaussian noise (AWGN) with two-sided power spectral density of $\frac{N_0}{2}$.

The interference cancellation scheme for each subcarrier is shown in Fig. 5. The received signal is first down-converted, and lowpass-filtered. Then, the signal is passed through a bank of correlators. Among K correlator outputs, the signal with maximum strength is chosen and cancelled from the received signal. The resulting received signal is again passed through a bank of correlators, and the next strongest one is selected and cancelled from the received signal. This process is repeated until the weakest user signal is detected.

4 Successive Interference Cancellation

The successive interference cancellation (SIC) detector takes a serial approach to cancellation interference. Each stage of this detector decisions, regenerates, and cancels out one additional direct-sequence user from the received signal, so that the remaining users see less MAI in the next stage. The SIC scheme implements the following steps:

1. Detect the strongest signal
2. Data decision on the strongest signal
3. Regenerate an estimate of received signal for the strongest user
4. Cancel out the strongest signal from the total received signal
5. A modified received signal without the MAI caused by the strongest user
6. The process is repeated until the weakest signal is detected.

5 Simulation Results

In this section, some simulation results are presented for the different system parameters. In Fig. 6, BER of a multicarrier DS-CDMA system is shown in an AWGN and Rayleigh fading channel for the number of user $K=8$ and single carrier. It is shown that the error performance in an AWGN channel is better than that in a Rayleigh fading channel.

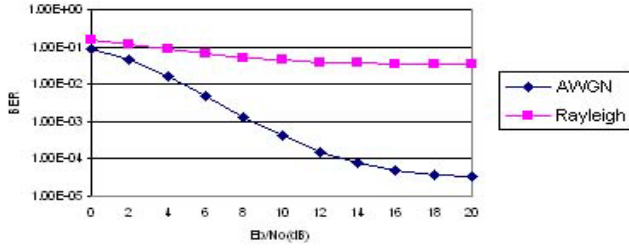


Fig. 6. BER versus E_b/N_0 in an AWGN and Rayleigh fading channel $M=1$ (single carrier) and $K=8$.

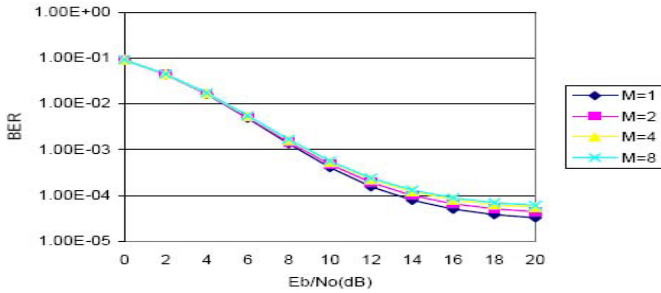


Fig. 7. BER without SIC versus E_b/N_0 for the various number of subcarriers in an AWGN channel, where $K=8$.

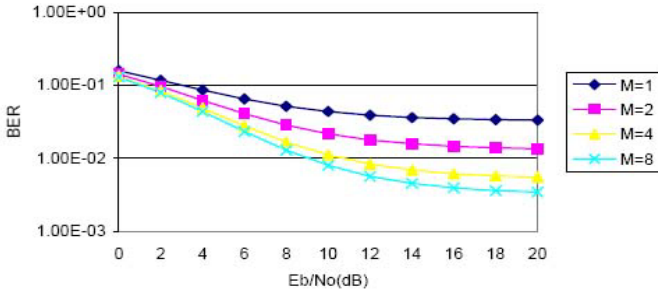


Fig. 8. BER without SIC versus E_b/N_0 for the various number of subcarriers in a Rayleigh fading channel, where $K=8$.

In Fig. 7, BER of a multicarrier DS-CDMA system without SIC is shown in an AWGN channel for the number of user $K=8$, and various number of subcarriers M . It is shown that the error performance without SIC is a little bit gradually degraded with the number of subcarriers in an AWGN channel.

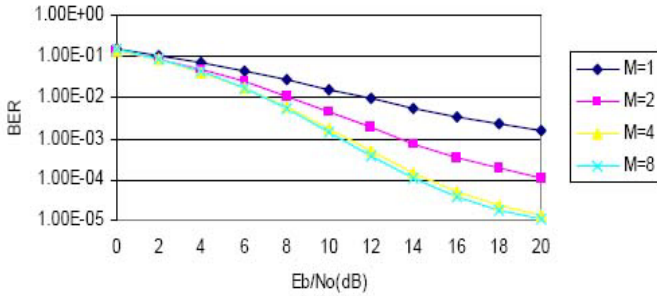


Fig. 9. BER with SIC versus E_b/N_0 for the various number of subcarriers in a Rayleigh fading channel, where $K=8$.

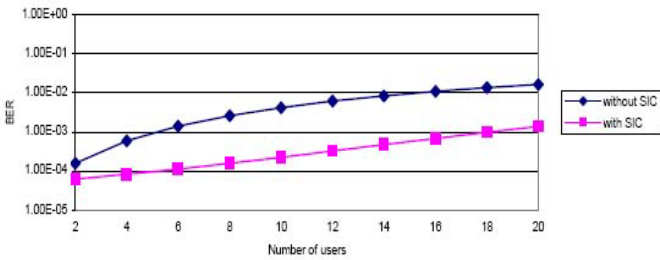


Fig. 10. BER versus number of users in a Rayleigh fading channel. $M=4$ and $E_b/N_0 = 15$ dB.

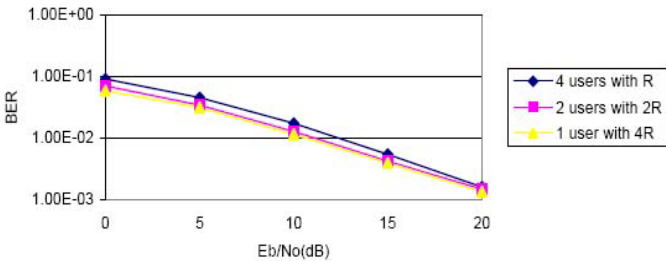


Fig. 11. BER with SIC versus E_b/N_0 for the different data-rate users in a Rayleigh fading channel.

In Fig. 8, BER of a multicarrier DS-CDMA system without SIC is shown in a Rayleigh fading channel for the number of user $K=8$, and various number of subcarriers M . As the number of subcarrier increases, the error performance becomes better.

In Fig. 9, BER of a multicarrier DS-CDMA system with SIC is shown in a Rayleigh fading channel for the number of user $K=8$, and various number

of subcarriers M . As the number of subcarrier increases, the error performance becomes better. In comparison to Fig. 8, the system with SIC achieves significant performance enhancement over the system without SIC.

Fig. 10. BER versus number of users in a Rayleigh fading channel. $M=4$ and $E_b/N_o = 15\text{dB}$. In Fig. 10, BER with and without SIC vs. the number of users is shown for $M=4$ and the received $E_b/N_o = 15\text{dB}$. As the number of users increases, BER with SIC increases much slower than without SIC. In Fig. 11, BER vs. E_b/N_o is shown for the three different data rate cases: 1) 4 users with R , 2) 2 users with $2R$, 3) 1 user with $4R$. As SNR increases, the performance difference among the users of different data rate decreases because the effect of MAI becomes less dominant as the SNR increase.

6 Conclusions

The SIC scheme was proposed and analyzed for performance enhancement of a multirate multicarrier DS-CDMA system. It was shown that, in terms of BER and system capacity, the multirate multicarrier DS-CDMA system with SIC achieves significant performance enhancement over the conventional DS-CDMA system. As the number of subcarriers increases, error performance of the considered system was gradually improved, but there is a limit on achievable improvement because cancellation error increases. It is also shown that one high-rate user achieves better performance than the other two schemes since one high data-rate user experiences no MAI. As SNR increases, the performance difference among the users of different data rate decreases because the effect of MAI becomes less dominant as the SNR increase.

Acknowledgments. This work has been supported by the Brain Korea 21 project in 2003, and in part, by Kwangwoon university in 2003.

References

1. W. C. Y. Lee, "Overview of cellular CDMA," *IEEE Trans. Veh. Technol.*, vol. 40, no. 2, pp. 291-302, May 1991.
2. A. Duel-Hallen, J. Holtzman, and Z. Zvonar, "Multi-User detection for CDMA systems," *IEEE Pers. Commun.*, vol. 2, pp. 46-58, Apr. 1995.
3. S. Kondo and L. B. Milstein, "On the use of multicarrier direct sequence spread spectrum systems," in *Proc. of IEEE MILCOM*, pp. 52-56, Boston, MA, Oct. 1993.

SCORM-Based Contents Collecting Using Mobile Agent in M-learning

Sun-Gwan Han¹, Hee-Seop Han², and Jae-Bong Kim¹

¹ Dept. of Computer Education, Gyeong-In National University of Education
Gyo-dae Street, Gye-yang-gu, Inchon, Korea, 407-753

² Department Of Computer Science Education, College of Education, Anam-dong
Sungbuk-ku, Seoul, 136-701, Korea
han@gin.ac.kr, anemon@korea.com, tudul@hanmail.net

Abstract. This study proposed the system for collecting the learning contents using the mobile agent in e-learning and m-learning. The mobile agent in wireless environment can provide the flexible works through a network and collect the adequate learning contents according to the learner's request. Moreover, the mobile agent can reduce the communication expenses. In this study, we proposed the CAS system that the mobile agent intellectually collected the appropriate learning contents in wireless learning. We used the SCORM-based contents as learning materials. In order to examine the validity of proposed study, we designed and implemented proposed system with mobile agent.

1 Introduction

The Internet-based distance learning and e-Learning have been changed the entire paradigm of an education. Now, the learner can access an education anywhere and whenever, and learn with a learner-centered method. These changes are demanded the advance of information and communication technology. This environment requests the inductive method of teaching to the teacher as well as needs the new content types as the learning materials. At these points, a wireless communication technology is accelerated such changes and environment. However, we are much room for consideration about the economical efficiency of the wireless communication. The wireless communication service has a policy of very high rate. We spend much time a searching data, a downloading data, and a verifying data in wireless environment. As a result, we must pay very expensive fare to get the contents and to use the mobile services. This disadvantage makes a burden to a learner and a teacher in learning and training. Some problems also appear as followings.

- A difficulty of searching by the increase of learning contents.
- A burden of communication expenses by the long connection time.
- The inexactness of the searched result by no certified learning contents.
- An additional expenses by downloading an unsuitable learning content.

To solve these problems, we proposed the content collecting system using a mobile agent and SCORM(Sharable Content Object Reference Model) in wireless

communication environment. Proposed system can provide the learner an effective m-learning through the characteristic of mobile agent as like mobility, autonomous and intelligent ability. That is, this system can service the learner the suitable learning contents, because the content is designed the application of SCORM. Moreover, we can decrease the communication expenses used by mobile agent. From next chapter, we show the method of design and implementation of the proposed system to collect the learning contents in m-learning.

2 Backgrounds

2.1 Mobile Communication Environment

Generally, an account policy of mobile communication expenses divides into the packet-based and the circuit-based system. The packet system is fixed price by the number of packets. The circuit system is fixed price by the connection time named DOSU in Korea. One packet is 128bytes, and one DOSU is 10 seconds. Mobile communication companies in current Korea, the fare of mobile communication are followings. In the type of circuit system, a charge for 10 minutes is 0.66 US\$. That is 60 DOSU (10 minutes) 0.011 US\$ = 0.66. In the type of packet system, one text (50 packets) is 0.25 US\$—0.005US\$/packet—, and small multimedia (100 packets) is 0.2US\$ (1 packet = 0.002US\$)[11]. Therefore, this account policy of mobile communication has been made the burden to the user.

2.2 Mobile Agent

The agent is divided the two types by the mobile ability. A stationary agent is a fixed program in the systems and is executed a task by transmitting the messages, while a mobile agent is a cloned program and is run in various systems after moving through the network. If the stationary agent communicates with other agents, the multi-agents use the communication mechanism such as Remote Procedure Calling(RPC). The other side, the mobile agent is not tied in system where the agent is executed. And the mobile agent can move a cloned agent code to other systems through network freely like as figure 1[3].

Therefore, the mobile agent system does not need more sustaining the connection of communication until an agent finds a given solution. As a result the mobile agent has the advantages such as the reducing of network overload, the overcoming of network latency, the encapsulating protocols, an asynchronous and autonomous execution, the adapting and dynamical reacts, the robust and fault-tolerant characters [2].

Mobile Agent is the agent that executes given tasks after arrival at the several destination systems through network. If we would use a mobile agent in m-learning, we can get the advantages of cost reduction like followings:

- The decreasing of searching time by the parallel and distribution works.
- In case of circuit system, the decreasing of connection time by an asynchronous work during data searching.

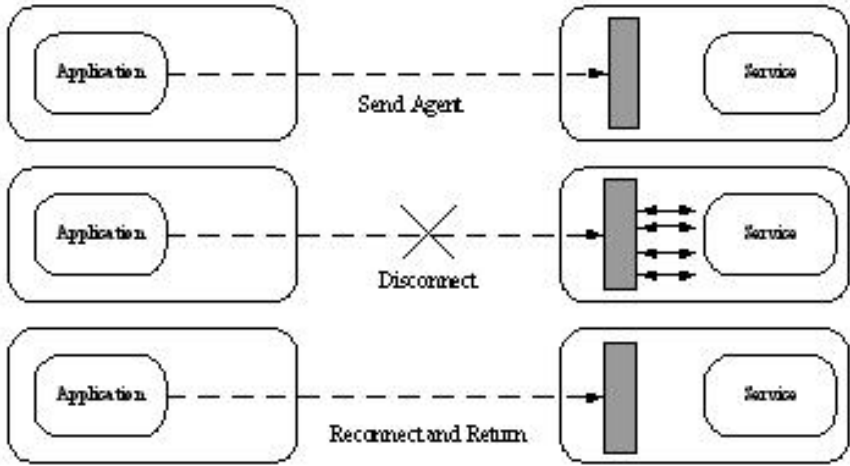


Fig. 1. Characteristic of Mobile Agent

- In case of packet system, the decreasing of packet transmission by the satisfaction of the searched data.

2.3 SCORM Overview

The SCORM by suggested ADL(Advanced Distributed Learning) is a reference model for standardization of WEB-based education contents. ADL proposed the SCORM to increase the rate of the using the technology-based learning, and to supply economic efficiency. The SCORM can make user use and modify freely the educational contents, regardless of the underlying hardware or operating system. SCORM defines the high-level requirements such as reuse, access and inter-operation for educational contents.

The SCORM defined the web-based Content Aggregation Model and Runtime Environment for learning objects. The purpose of Content Aggregation Model is to supply the common method, which can extract contents from resources (reusable, accessible, inter-operable) and organize it [9]. The SCORM has the merits of followings:

- Accessible:* Content can be identified and located when it is needed and as it is needed to meet training and education requirements.
- Durable:* Content does not require modification to operate as software systems and platforms are changed or upgraded.
- Inter-operable:* Content will function in multiple applications, environments and hardware and software configurations regardless of the tools used to create it and the platform on which it is delivered.

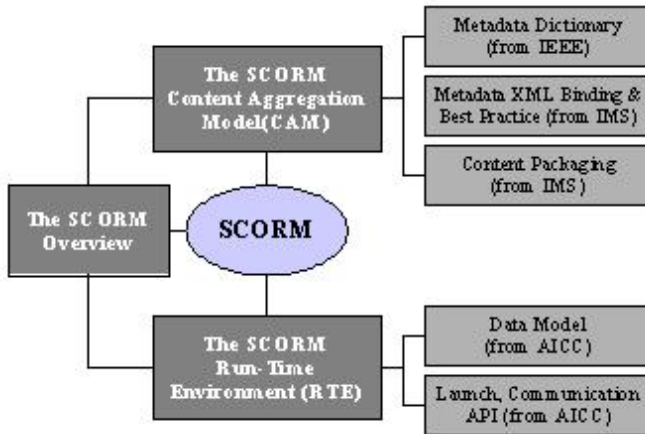


Fig. 2. SCORM Overview

Reusable: Content is independent of learning context. It can be used in numerous training situations or for many different learners with any number of development tools or delivery platforms.

Adaptable: by tailoring instruction to the individual and organizational needs. *Affordable:* by increasing learning efficiency and productivity while reducing time and costs.

3 Design of the CAS System

3.1 M-learning Environment

The wireless learning environment, which is proposed by this study, is the unity of the wireless and the wired like a following figure. The learner searches the wanted contents with mobile device-phone, PDA, notebook, and so on. CAS (Collecting Agent Server) creates the mobile agent and manages the collected contents. The mobile agent has the rules for executing a given-job. The SCORM-based learning contents exists in the LCMS(Learning Contents Management Systems). The mobile agent collects the suitable contents for a learner with the information of SCORM metadata. The learner's searching actual process is followings:

1. Query: A learner queries with a mobile phone, and connects at CAS, and transfer the query, and disconnects.
2. Create agent: CAS receives the query, and creates mobile agents.
3. Dispatch agent: CAS dispatch the mobile agent together with the query to LCMS or other learning contents systems.
4. Collect contents: The mobile agent searches and collects the information of suitable contents in the SCORM-based LCMS.

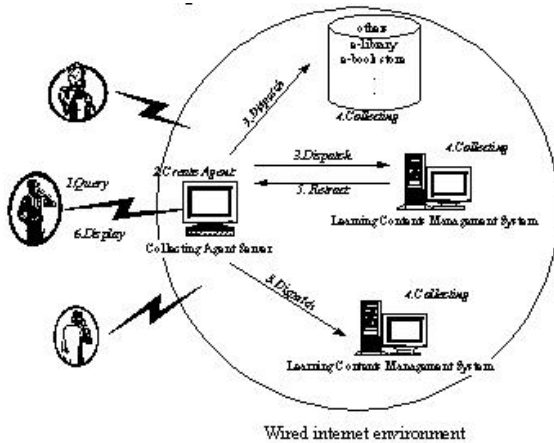


Fig. 3. Wireless Learning Environment

5. Retract agent: The mobile agent retracts to CAS with the contents? collected information.
6. Display result: CAS supplies the learner the more suitable results in the collected contents through wireless.

3.2 SCORM-Based Learning Content

The SCORM-based learning contents apply LOM(Learning Object Metadata), which is proposed by LTSC. The mobile agent collects the suitable contents with the metadata. The inference engine in CAS selects the more suitable results in the collected contents. And CAS supplies the learner the best contents. The agent infers with the metadata and the learner's profile in searching.

The Base Scheme consists of nine such categories: General, Lifecycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation, and Classification category. The metadata categories, which the mobile agent and the CAS use in such categories, are followings.

3.3 System Architecture

CAS Architecture: CAS mainly consists of Communication module, Inference Engine, Agent Management Module, Learner Management Module, and Contents Management Module. The Communication Module is in charge of the communication between the learner and LMS. It uses HTTP protocol with the learner and ATP (Agent Transfer Protocol) with LCMS. The Inference engine interacts an Agent Management Module, a Learner Management Module, and a Contents Management Module. Also, it selects the most suitable results. Agent

Table 1. Basic Scheme of categories in SCORM

Main category	Sub-Category	Main category	Sub-Category
General	Identifier, Title, Catalog Entry, Catalog Entry, Language, Description, Keyword, Coverage, Structure	Educational	Interactivity Type, Learning Resource Type, Interactivity Level, Semantic Density, Intended End User Role, Context, Typical Age Range, Difficulty, Typical Learning Time, Description, Language
Lifecycle	Version, Status, Contribute, Role, Entity, Date	Rights	Cost, Copyright and Other Restrictions, Description
Meta-meta data	Identifier, Catalog Entry, Catalog Entry, Contribute, Role, Entity, Date, Language	Relation	Kind, Resource, Identifier, Description, Catalog Entry, Catalog, Entry
Technical	Format, Size, Location, Type, Requirement, Name, Minimum-maximum Version, Installation Remarks, Other Platform Requirements, Duration	Annotation	Person, Date, Description
		Classification	Purpose, Tax on Path, Source, Tax on, Id, Entry, Description, Keyword

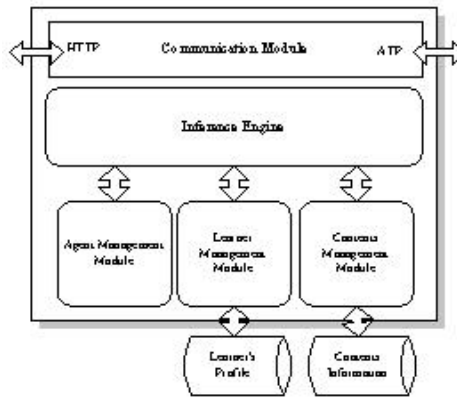


Fig. 4. CAS Architecture

Management Module creates and manages the agents. Learner Management Module manages the learner’s profile for better result. Contents Management Module manages the information of the contents, which the mobile agent brought.

LCMS Architecture: LCMS has the general learning contents, and connects LMS (Learning Management System). In this study, LCMS has the SCORM-based contents. LCMS has agent run-time environment. The agent run-time environment consists of Communication Module, Agent Place, and Contents Searching Module. Agent Place is a context in which an agent can execute. The agents interact with the Contents Searching Module. Contents Searching Module does search with the metadata, which embedded in contents.

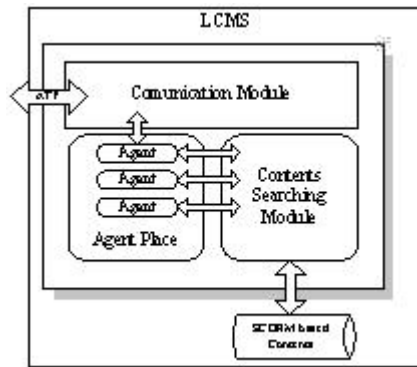


Fig. 5. LCMS Architecture

4 Implementation of the CAS System

The agent run-time environment is implemented Aglet API, which is developed by IBM Tokyo LAB in Japan. The Aglet API is an agent development kit - in other words, a set of JAVA classes and interfaces that allows you to create mobile Java agents. We set CAS in Linux Server, and implemented the intractable CAS with Aglet API. We made LCMS with MySQL Server in Linux Server. Content database connects the mobile agent with JDBC. We had regard to a notebook, and implemented the Interface Module. We implemented the User Interface, with which the learner could select the items of the educational category in the SCORM main metadata category and could search. When the given-keyword and the selected items are transferred in CAS, The Aglet executes through the Servlet. The User interface is followings.

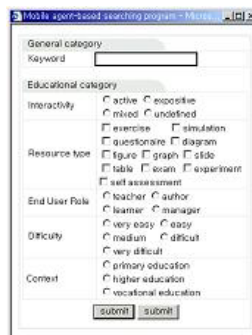


Fig. 6. Interface for User

5 Conclusions

This study proposed the method that collects the learning contents with the mobile agent effectively in the wireless learning environment. It can cut the cost of the communication expenses with the asynchrony of the mobile agent. Also, it can supply the learner the elaborate contents because of the intelligence of the mobile agent and the SCORM. This study display the example that is implemented in the limited environment such Aglet. However, we expect that this proposal will utilize effectively in the wireless environment, because of the standardization policy of the mobile agent in the future.

We need to make researches into the method, which is the mobile agent to search effectively the learning contents. We need to research the learning packaging and the learning sequencing to supply the suitable learning with the rule-based inference engine of the SCORM metadata and the learner's profile. We have to develop the run-time environment, which can apply in PDA and mobile phone.

References

- [1] Buckley, C., Implementation of the SMART Information Retrieval System, Technical Report 85-686, Cornell University, (1985)
- [2] Chang, D., Lange, D. Mobile Agent: A new paradigm for distributed object computing on the WWW. In Proceedings of the OOPSLA96 Workshop: Toward the Integration of WWW and Distributed Object Technology, ACM Press, N.Y., (1996), 25-32.
- [3] Danny D. Lange, Mitsuru Oshima, Programming and Deploying Java Mobile Agents with Aglets, Addison-Wesley, (1998)
- [4] Dublin Core Metadata Initiative. <http://www.dublincore.org/>.
- [5] IEEE Information Technology - Learning Technology - Learning Objects Metadata LOM: Working Draft 6.1 (2001-04-18). As referenced by the IMS Learning Resource Meta-data Specification Version 1.2.<http://ltsc.ieee.org/>. (2001)
- [6] Institute of Electrical and Electronics Engineers (IEEE) Learning Technology Standards Committee (LTSC), (2003)
- [7] J.White, Mobile Agents, in Software Agents, J. Bradshaw, ed., MIT Press, Cambridge, MA, (1997)
- [8] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, (1999)
- [9] Sharable Content Object Reference Model(SCORM). Version 1.2. ADL, 2001-11-14. (<http://www.adlnet.org//library/documents/scorm/software/SCORM.htm>), (2001)
- [10] Sun-Gwan Han, Young-GI Kim, Jae-Bok Park, 'Cooperative Monitoring System using Mobile Agent', ICCE2000, Taiwan, (2000)
- [11] Wireless Internet Corporation in Korea, <http://www.nate.com>, (2003)

Improved Bit-by-Bit Binary Tree Algorithm in Ubiquitous ID System

Ho-Seung Choi, Jae-Ryong Cha, and Jae-Hyun Kim

School of Electrical and Computer Engineering,
AJOU University, San 5 Wonchon-Dong, Youngtong-Gu, Suwon 442-749, Korea
{lastjoin, builder, jkim}@ajou.ac.kr

Abstract. We propose and analyze the fast wireless anti-collision algorithm(Improved Bit-by-bit Binary-Tree algorithm(IBBT)) in Ubiquitous ID system. We mathematically compare the performance of the proposed algorithm with that of Binary Search algorithm(BS), Slotted Binary-Tree algorithm(SBT) using time slot, and Bit-by-bit Binary-Tree algorithm(BBT). We also validated analytic results using simulation. According to the analytical result, comparing IBBT with BBT which is the best of existing algorithms, the performance of IBBT is about 304% higher when the number of the tags is 20, and 839% higher for 200 tags.

1 Introduction

In a world of ubiquitous computing, the object identification is most useful for applications such as asset tracking(e.g. libraries, animals), automated inventory and stock-keeping, toll collecting, and similar tasks where physical objects are involved. The RFID(Radio Frequency Identification) systems are a simple form of ubiquitous sensor networks that are used to identify physical objects. In this paper, we call RFID system as Ubiquitous ID(u -ID) system. Instead of sensing environmental conditions, the u -ID system identifies the unique tags' ID or detailed information saved in them attached to objects. Passive u -ID systems generally consist of three components - a reader, passive tags, and a controller. A reader interrogates tags for their ID or detailed information through an RF communication link, and contains internal storage, processing power, and so on. Tags get processing power through RF communication link from the reader and use this energy to power any on-tag computations and to communicate to the reader. A reader in u -ID system broadcasts the request message to the tags. Upon receiving the message, all tags send the response back to the reader. If only one tag responds, the reader receives just one response. If there are more than one tag response, their responses will collide on the RF communication channel, and thus cannot be received by the reader[1]. The problem of resolving this collision is referred to as the Anti-Collision Problem, and the ability to resolve this problem is crucial in u -ID system performance. In u -ID system, important measures of performance include the time required to identify the tags, and the power consumed by the tags. If the data from the tags are small, we

can reduce the time to identify the tags and the power consumed by the tags. This paper mathematically compares the performance of the proposed algorithm with that of Binary Search algorithm(BS), Slotted Binary-Tree algorithm(SBT), and Bit-by-bit Binary Tree algorithm(BBT). We also validate analytical results using OPNET simulation.

2 Existing Binary Anti-collision Algorithms

2.1 Binary Search Algorithm(BS) [2]

BS resolves the collision by reducing gradually collided bit in a tag ID. When a collision is occurred, the reader knows the position in which the collision occurs. If there are tags whose first collided bit is 1, they do not respond to the reader's next request. But tags whose first collided bit is 0 transfer their ID to the reader's next request. By repeating this procedure, the reader can identify all the tags. More details are in [2]. Assuming that there are tags, the number of iterations of BS(I_{BS}) is

$$I_{BS} = \frac{\log(n)}{\log(2)} + 1. \quad (1)$$

2.2 Slotted Binary Tree Algorithm(SBT) [3]

According to this algorithm when a collision occurs, in slot i , all tags that are not involved in the collision wait until the collision is resolved. The tags involved in the collision split randomly into two groups, by(for instance) each selecting 0 or 1. The tags in the first group, those that selected 0, retransmit in slot $i+1$ while those that selected 1 wait until all tags that selected 0 successfully transmit their ID. If slot $i+1$ is either idle or contains a successful transmission, the tags of the second group, those that selected 1, retransmit in slot $i+2$. If slot $i+1$ contains another collision, the procedure is repeated[3]. Assuming that there are n tags, the number of iterations of SBT(I_{SBT}) is

$$I_{SBT} = 1 + \sum_{k=2}^n \binom{n}{k} \frac{2(k-1)(-1)^k}{[1-p^k - (1-p)^k]}. \quad (2)$$

2.3 Bit-by-Bit Binary Tree Algorithm(BBT) [4],[5]

In BBT, if the reader requests ID bit to the tags, all the tags transfer 0 or 1 out of their ID bit. If not colliding, the reader saves the received bit in the memory before it requests next bit. But if colliding, the reader selects one group out of two groups according to the algorithm, then requests next bit. The procedure is as follows. If the reader requests k^{th} bit from the tags, the tags transfer k^{th} bit.(In this algorithm, the initial value of k is 1). If the reader receives k^{th} bit without collision, it saves k^{th} bit in the memory. But if colliding, the reader saves k^{th} bit as 0 in the memory. The reader then makes all the tags whose k^{th} bit is

Table 1. Used tags' ID

Tag1	0001
Tag2	0010
Tag3	1010
Tag4	1011

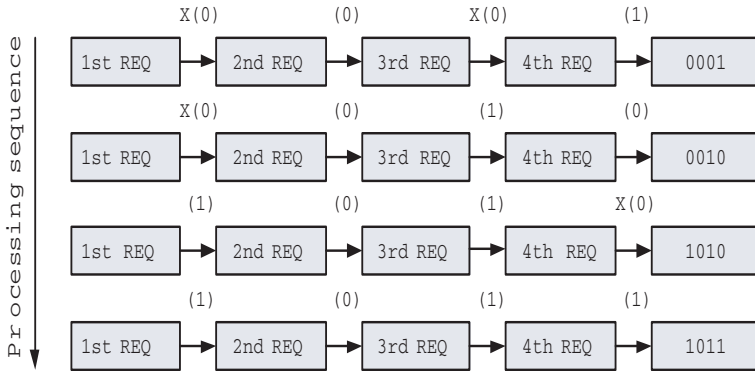


Fig. 1. Bit-by-bit binary-tree algorithm

1 inactive. The tags in inactive state do not temporarily transfer their ID bits until one tag is identified. The tags in inactive state do not temporarily transfer their ID bits until one tag is identified. The reader repeats this procedure until all the bits of an ID are received. Assuming that there are n tags and the length of tag ID is j bits, then the number of iterations of BBT(I_{BBT}) is

$$I_{BBT} = n \times j. \tag{3}$$

Fig.1 shows the process to identify four tags in Table 1. In Fig.1, 'X' means collision and the number of total transferred bits from the tags to identify four tags is $16(4 \times 4)$ bits).

3 The Improved Bit-by-Bit Binary-Tree Algorithm(IBBT)

In BBT, the reader always requests all the bits of tags' ID, so it takes much time to identify all the tags. To resolve this problem, we propose IBBT. IBBT procedure is as follows. First of all, the reader requests all the bits of tags' ID. When the reader receives k^{th} bit of tags' ID and if all the bits are 0(1), the reader saves k^{th} bit as 0(1). But if the collision occurs, the reader saves k^{th} bit as X in the memory. After the reader receives all the bits of tags' ID, the reader knows which bits are collided. The reader sequentially re-requests only collided bits to the tags by the method of bit-by-bit. If there are two tags whose ID except for

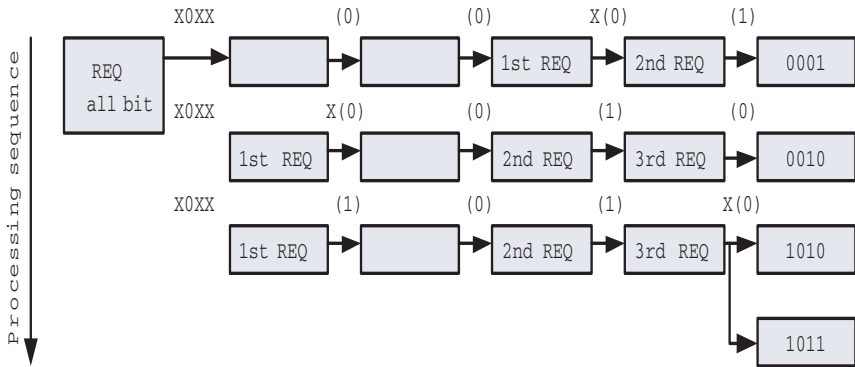


Fig. 2. The proposed IBBT algorithm

the last bit is identical, the last bit transferred from tags will definitely collide. In this case, we do not need to request next bit in IBBT. Accordingly, we can identify two tags simultaneously. As a result, the number of iterations of IBBT is less than that of BBT. If the tags' ID is sequential, the number of collided bits reduces. Therefore, the performance of IBBT is better than that of BBT.

Fig.2 shows the process to identify four tags in Table 1 using IBBT. In this example two tags, whose IDs are '1010' and '1011' respectively, have identical ID except the last bit. When the reader requests these tags' last bit, the collision occurs. However the reader can identify two tags simultaneously without further request. In Fig.2, the number of total transferred bits from the tags to identify four tags is 12.

4 Performance Analysis

In this section, we analyze the performance of IBBT from two points of view. One is the number of iterations and the other is total transferred bits from the tags.

4.1 The Number of Iterations

We assume that arbitrary m tags are used out of n tags whose ID is sequential, and the length of ID is 36 bits[5]. Let II be the number of iterations which does not consider the tags whose last ID bit collides. Then II is driven by (4).

$$\begin{aligned}
 II = & \left(\frac{2^r}{n} + m + 1 \right) (r + 1) + \frac{m \left(1 - \frac{2^r}{n} \right)}{2} r + \frac{m \left(1 - \frac{2^r}{n} \right)}{4} (r - 1) + \dots \\
 & + \frac{m \left(1 - \frac{2^r}{n} \right)}{2^{k_{\max}}} (r + 1 - k_{\max}) + \left(\frac{m \left(1 - \frac{2^r}{n} \right)}{2^{k_{\max}}} - 1 \right) (r - k_{\max})
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{\log_2 m \left(1 - \frac{2^r}{n}\right)} \frac{m \left(1 - \frac{2^r}{n}\right)}{2^k} (r+1-k) + \left(\frac{2^r}{n} \times m + 1\right) (r+1) \\
&\quad + \left(\frac{m \left(1 - \frac{2^r}{n}\right)}{2^{k_{\max}}} - 1\right) (r - k_{\max}) \tag{4}
\end{aligned}$$

When $k_{\max} = \lceil \log_2 m \left(1 - \frac{2^r}{n}\right) \rceil$, $2^r < n \leq 2^{(r+1)}$, and $0 < r \leq 35$ (r is a integer). And $\lceil * \rceil$ is a maximum integer less than or equal to $*$. We can calculate the number of iterations of IBBT by multiply Π by the term which considers the tags whose last ID bit is collided. When the number of used tags(m) is even, the number of iterations of IBBT(I_{EBBT}) is calculated by (5).

$$I_{EBBT} = \begin{cases} \Pi \times \frac{1}{m} \sum_{k=0}^{\frac{m}{2}} \frac{\binom{\frac{n}{2}}{k} \binom{\frac{n}{2}-k}{m-2k} 2^{(m-2k)}}{\binom{n}{m}} \cdot (m-k), & 0 < m \leq \frac{n}{2} \\ \Pi \times \frac{1}{m} \sum_{k=m-\frac{n}{2}}^{\frac{m}{2}} \frac{\binom{\frac{n}{2}}{k} \binom{\frac{n}{2}-k}{k-m+\frac{n}{2}} 2^{(m-2k)}}{\binom{n}{m}} \cdot (m-k), & \frac{n}{2} < m \leq n \end{cases} \tag{5}$$

Otherwise, the number of used tags(m) is odd, the number of iterations of IBBT(I_{IBBT}) is calculated by (6).

$$I_{IBBT} = \begin{cases} \Pi \times \frac{1}{m} \sum_{k=0}^{\frac{m-1}{2}} \frac{\binom{\frac{n}{2}}{k} \binom{\frac{n}{2}-k}{m-2k} 2^{(m-2k)}}{\binom{n}{m}} \cdot (m-k), & 0 < m < \frac{n}{2} \\ \Pi \times \frac{1}{m} \sum_{k=m-\frac{n}{2}}^{\frac{m-1}{2}} \frac{\binom{\frac{n}{2}}{k} \binom{\frac{n}{2}-k}{k-m+\frac{n}{2}} 2^{(m-2k)}}{\binom{n}{m}} \cdot (m-k), & \frac{n}{2} < m < n \end{cases} \tag{6}$$

4.2 Total Transferred Bits from the Tags

Let I be the number of iterations for each algorithm and B_I be the transferred bits from the tags in each iteration, then the number of total transferred bits(B_{total}) from the tags is

$$B_{total} = I \cdot B_I. \tag{7}$$

In IBBT, when the reader initially requests all ID bits from the tags, the tags send all ID bits to the reader. So, the number of total transferred bits from the tags should be added by 35(the length of ID-1).

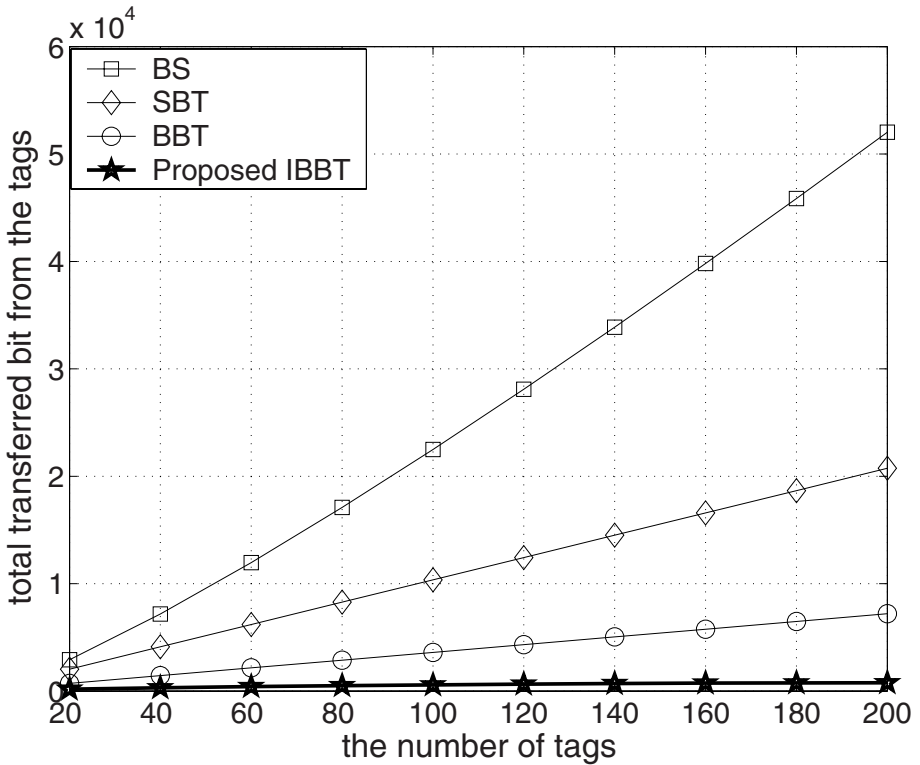


Fig. 3. Total transferred bits from the tags vs. the number of the tags

5 Numerical and Simulation Results

Fig.3 shows total transferred bits from the tags for the number of tags in each algorithm and represents the analytic results. In proposed algorithm, all tags transfer one bit whenever the reader requests ID from the tags. But all other algorithms except for BBT transfer all the bits of tags' ID. So, the proposed algorithm reduces the time to identify tag and the energy consumed by the tag, because the number of transferred bits from the tags is smaller compared to the existing algorithms. In Fig.3, to identify 100 tags, BS, SBT, and BBT transfer 22491, 9482, and 3600 bits respectively. For the same scenario, IBBT only transfers 578 bits. Therefore, we found that proposed algorithm has very higher performance compared to the existing algorithms.

Fig.4 shows the number of iterations for the number of tags in BBT and IBBT according to the number of tag's ID bit[6]. In Fig.4, lines represent analytic results and symbols represent simulation results using OPNET. Analytic results are very closed to the simulation results. Comparing IBBT with BBT, the number of iterations in BBT increases as the number of tag's ID bit increases, while that of IBBT does not increase according to the number of tag's ID bit.

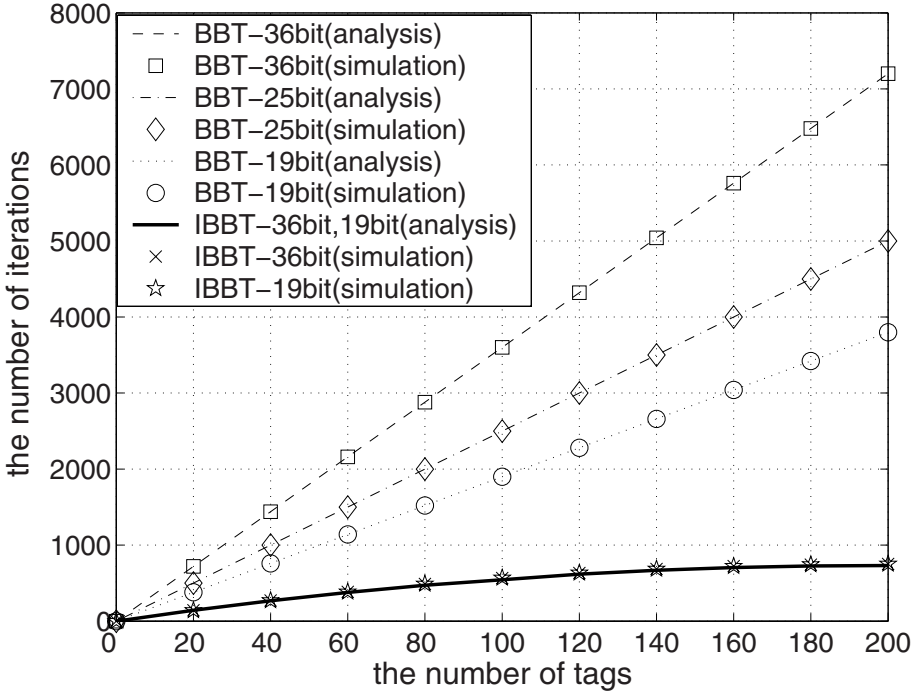


Fig. 4. The number of iterations vs. the number of tags

That is why the IBBT shows much better performance for the larger tag's ID bit. We found that the performance of IBBT is 304% higher when the number of tags is 20 for the 36 bit tag's ID. Moreover, we found that the performance of IBBT is about 839% higher when the number of tags is 200. As the number of tags increases the performance of IBBT is much better than that of exiting algorithms.

6 Conclusion

We proposed and analyzed the fast wireless anti-collision algorithm (IBBT) in u -ID system. We mathematically compared the performance of the proposed algorithm with that of BS, SBT, and BBT. We also validated analytic results using simulation. IBBT shows better performance than existing algorithms as the number of tag's ID bit increases. The energy consumption is also lower than existing algorithms since the number of bits transferred from the tags is much smaller. Furthermore, the performance of IBBT is much better than exiting algorithms as the number of tags increases. On conclusion, if we apply the proposed algorithm to the u -ID system, it will contribute to improve the performance of the u -ID system because the reader can identify more tags with shorter time and less energy.

Acknowledgment. This research is partially supported by the Ubiquitous Autonomous Computing and Network Project, the Ministry of Science and Technology(MOST) 21st Century Frontier R&D Program in Korea.

References

- [1] H. Vogt, 'Efficient Object Identification with Passive RFID tags,' In International Conference on Pervasive Computing, Zurich, (2002).
- [2] K. Finkenzeller, RFID Handbook: *Radio-Frequency Identification Fundamentals and applications*. John Wiley and Sons Ltd, (1999).
- [3] J. L. Massey, 'Collision resolution algorithms and random-access communications,' Univ. California, Los Angeles, Tech. Rep. UCLAENG -8016, Apr. (1980).
- [4] M. Jaoccurt, A. Ehram, U. Gehrig, 'Contactless Identification Device With Anticollision Algorithm,' IEEE Computer Society CSCC'99, Conference on Circuits, Systems, Computers and Communications, Jul. 4-8 Athens (1999).
- [5] Auto-ID Center, *Draft Protocol Specification for a Class 0 Radio Frequency Identification tag.*, (2003).
- [6] EPC Global, *EPCTM Tag Data Standards Version 1.1 Rev.1.24*, Apr. (2004).

Automatic Synchronized Browsing of Images Across Multiple Devices

Zhigang Hua^{1*}, Xing Xie², Hanqing Lu¹, and Wei-Ying Ma²

¹ National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100080, P.R.China
{zghua, luhq}@nlpr.ia.ac.cn

² Microsoft Research Asia, No.49, Zhichun Road, Beijing, 100080, P.R.China
{xingx, wyma}@microsoft.com

Abstract. Mobile devices are undergoing considerable progress during recent years. Using these portable devices, people can easily capture and share photos even when they are on the move. Doubtlessly, browsing a large number of images on such small-form-factor devices is still hard and time-consuming, especially when they are distributed across various devices. In this paper, we propose a novel synchronized approach to facilitate image browsing across multiple devices. In this approach, similar images across multiple devices can be simultaneously presented for users to make comparatively viewing or searching. Experimental results show that the synchronized approach is beneficial to improve users' browsing experience.

1 Introduction

Mobile devices are undergoing considerable progress in both hardware and software development during recent years. Using these portable devices, people can easily capture and share images even when they are on the move. For example, a recent popular trend is moblogging (mobile weblogging). It is through the use of a phone or other mobile devices for users to publish and share their resources on the Web in real time, whether that resource is text, image or other media. However, to make people really enjoy the ease of mobile communications, many hurdles still need to be crossed [4]. Among them, major crucial challenges include the limited accessing bandwidth and display sizes. While the bandwidth condition is expected to be greatly improved within the following years, however, in the foreseeable future, the display, i.e. the form factor, will continue to be the major constraint on small mobile devices.

To deal with the display constraint, attention model [1,3] have been proposed recently to facilitate image browsing in small devices. In [1], it allows the delivery of more important regions instead of whole images to clients when their screen sizes are small. In [3], an image is decomposed into a set of spatial-temporal

* This work was performed when the first author was working as a visiting student at Microsoft Research Asia.

information elements which are displayed serially, each for a brief period of time. Although these approaches proved to be effective for browsing large images on devices with small displays, searching or comparing a large number of images across multiple mobile devices is still a hard task.

In this paper, we propose a synchronized approach to facilitate image browsing across multiple devices. In our approach, similar images on various devices can be simultaneously presented for comparatively viewing or searching. The ability to view multiple similar images across devices at one time is useful, such as when users want to make a comparison or search through image collections from various devices. In the novel approach, traditional image browsing interface constrained to one single device is extended to multiple devices. The rest of this paper is organized as follows. Section 2 introduces the system framework of our approach. Section 3 discusses in detail the communication protocol that is adopted in our synchronized approach. In Section 4, the image matching algorithm is presented. In Section 5, we give the experimental results in our use study. Finally, concluding remarks are described in Section 6.

2 Our System Framework

The details of our system framework will be described in this section, they are, user interface we adopt in use, the synchronization across devices implicitly indicated by implicit query of attention objects, and the system flow of our framework.

2.1 User Interface

An image attention model [1,3] is generated to reveal the informative regions in images as:

$$AO_i = \{ROI_i, AV_i, MPS_i, MPT_i\} \quad (1)$$

In addition to all previous automatic image browsing approaches making use of the image attention model [1,3], however, in this paper we provide a smart browsing mode named “smart navigation” to implement the image browsing. In this mode, pressing the directional buttons in mobile devices will result in a smooth scrolling to the closest attention object in the given direction. Figure 1 (a) shows an example for this mode where the current focused attention object can be navigated to three closest objects by pressing left/right/up buttons. By using such a smart navigation, it can facilitate image browsing on small-form-factor devices with a substantive reduction of interactions. In our approach, synchronized image browsing across devices is proposed according to the attention object that displays on the screen, which is assumed to represent a user’s interest in the image. That’s to say, the attention object that currently displays is used as an implicit query to maintain synchronization on other devices, by searching through their image libraries to find out images containing similar attention objects. Furthermore, a function is offered by us to facilitate the viewing of these search images, by automatically zooming into the matched objects with

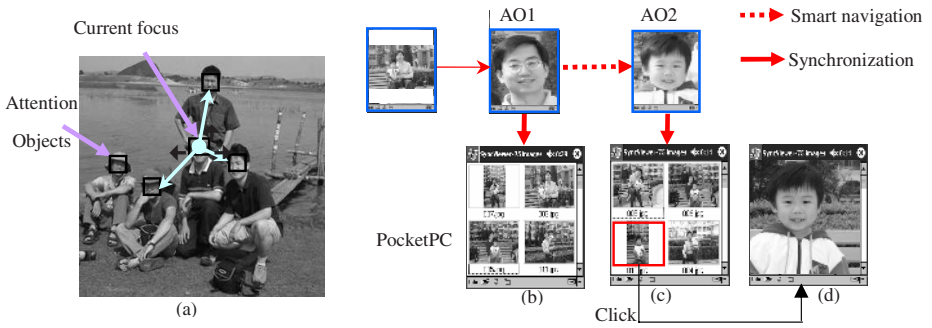


Fig. 1. (a) Smart navigation. (b) Synchronization results of AO1. (c) Synchronization results of AO2. (d) Automatic zooming into the matched attention object.

an appropriate display area when users click them (shown in Figure 1 (d)), which is to be described in details later.

2.2 Synchronization Based on Implicit Query

Since the synchronization across devices is implicitly indicated by the attention object that displays on the screen, we call this implicit query compared with explicit query where query task is directly specified by users such as web search engines. A synchronized process across devices can be summarized like this (in Figure 1, (b) and (c) respectively shows the synchronization results of two attention objects denoted AO1 and AO2):

1. A user selects one as a master device from multiple available devices, and begins to specify an image to browse through its attention objects using the smart browsing mode.
2. The feature of the current attention object is automatically extracted, and it is then delivered to the slave devices.
3. When a slave device receives such data containing feature of an attention object, it automatically starts to search through local image library to find out images containing similar features with the received one. The search images are sorted in an order according to the similarity and are then displayed on the screen.
4. If the user clicks an image in the search results on a slave device, the window automatically zooms into the matched attention object with appropriate display area.
5. Repeat steps 1 to 4 for images to be browsed across devices.

2.3 System Flow

In our synchronized approach, the feature of an attention object that currently displays on the master device is identified first, and is then automatically transferred to the slave devices for them to search for similar photos. When a user

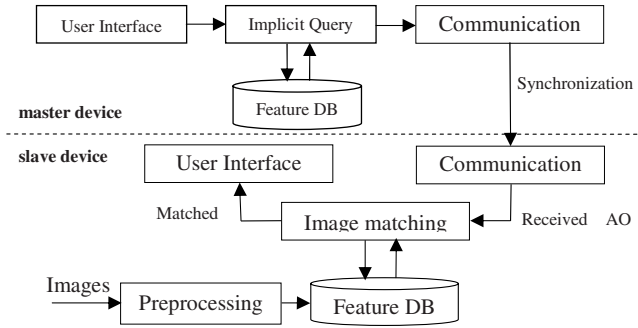


Fig. 2. The framework of synchronized image browsing.

interacts with a master device’s image using the smart browsing mode, the synchronized updates on other devices will be generated concurrently to display similar images. A complete system flow of our approach is shown in Figure 2. As shown in Figure 2, the framework mainly consists of three function modules:

1. The communication module is responsible for maintaining the synchronization relationship across various devices, that is, delivering feature data from a master device to its slave devices.
2. The image matching module is in charge of searching through local image library to find out images containing similar features with the received one.
3. The image preprocessing module is to save attention models and image features of each attention object as metadata into a local feature database for reuse. This is mainly due to the thin computing power in mobile devices such that the processing cost can be saved in real time.

3 Communication Protocol

To make the synchronized image browsing approach work effectively, the communication module plays an important role in maintaining the synchronization relationship across mobile devices. In the protocol, the devices are classified into two categories according to their usages or functions: master device is defined as the device that is operated by users; slave device is defined as the device which works synchronously with the master device. Several mobile devices can form a synchronization group by using wireless connection techniques like Bluetooth or 802.11b. Each group can only have only one master device at a time. Instead of directly delivering the whole image content, communication module only transmits the feature of an attention object, which is defined as:

$$F_{trans} = \{AO, Type, Feature\} \tag{2}$$

In the equation, three properties are assigned to the attention object to be transmitted. F_{trans} represents the feature data of an attention object to be delivered from a master device to its slave devices. AO stands for the properties

of an attention object, which is previously described in Equation 1. Type is to indicate the category an attention object, which is classified into saliency, face and text by [1]. This acts as a matching criterion in the image matching module discussed later. Feature is the low-level feature extracted from the current attention object, such as color, texture, moment, etc.

4 Image Matching

To achieve synchronized image browsing across devices, it is necessary for us to develop a method to search out synchronization images from an image collection with a large amount on various devices. In this section, our method is described in three aspects including feature extraction from images, image matching algorithm and a function to facilitate the viewing of synchronization images.

Features to Maintain Synchronization. In order to save the computational cost consumed in image processing, the feature extraction from images is done in the preprocessing module. In our approach, images are first analyzed to extract attention objects using the attention model proposed in [1]. In addition to the properties defined in Equation 1, in this paper, we extract low level image feature from each attention object for the use in the image matching. Many kinds of low-level features can be extracted from an attention object, however, we adopt color moment as the feature in our current implementation, which is shown to be robust and effective [2]. We extract the first two moments from each channel of CIE-LUV color space. In future work, we are considering using some semantic image features such as location, time, subject and event to extend the low-level feature that is currently used.

Image Matching Algorithm. More detailedly, assume there is one attention object AO_o within an image I_0 being browsed on the master device. Suppose an image collection in a slave device be $I = \{I_1, I_2, \dots, I_N\}$ with N images, and I_i ($1 \leq i \leq N$) is defined as the set of its included attention objects: $I_i = \{AO_{ik} | 1 \leq k \leq N\}$ (where n_i is the number of objects in I_i). In our implementation, the similarity between two attention objects is measured using the Euclidean distance. Figure 3 gives the process to make image matching. Here is the step-by-step description of the algorithm to find out similar objects with AO_o from I :

1. For each image I_i ($1 \leq i \leq N$) within the image collection I , do the step 2.
2. Examine the type of each attention object AO_{ik} ($1 \leq k \leq n_i$) within I_i . Group those attention objects containing the same type with AO_o into a set, let it be I'_i (which includes n'_i members, and it satisfies: $1 \leq n'_i \leq n_i$). Do the steps 3-4.
3. Calculate the Euclidean distance d_{ik} between the low-level feature of AO_o and each attention object AO'_{ik} within I'_i . Select AO'_{it} ($1 \leq t \leq n'_i$) with the least distance d_{it} , which is used to measure the similarity between AO_o and I_i .

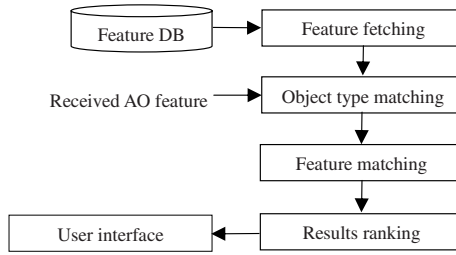


Fig. 3. The process to make image matching.

4. Identify whether the distance d_{it} is under an assigned threshold (r). If the distance is under this threshold, the corresponding AO'_{it} and d_{it} is inserted into a set denoted Q_0 .

By this way, Q_0 is gradually formed. The search results are finally displayed on the slave devices in an order according to the similarity calculated by the above steps.

Automatic Zooming into the Matched Attention Object. We provide a further function to facilitate image viewing on the search results. That is, when users specify one of images in the search results to click, the window will automatically zoom into the appropriate attention object using a display area no less than its property MPS (seen in Equation 1). This can be implemented by using the ROI property (seen in Equation 1), which represents the region information of an attention object within an image. The generation of automatic zooming into the matched attention object is referred to the method of the generation of automatic browsing path proposed in [3]. This function can let users view interesting regions without too much interactions like zooming/scrolling.

5 Experimental Results

Our evaluations were conducted with a Smartphone (Dopod 515) and a PDA (Compaq iPAQ 3670 Pocket PC). Each of them stores 75 images ranging from a variety of types including family, indoor, outdoor, group. We asked eight computer science graduate students to participate with our use study, including four males and four females. They are familiar with the operations in mobile devices like PDA and Smartphone, and they never have any knowledge of our synchronized approach before. The eight subjects were firstly asked to view an easy instruction on how to operate in our approach. After users get familiar with it, user evaluation was then performed to measure the performance.

Task-Based Evaluations. The participants were first asked to finish two tasks as follows: 1) Find out all images on the two devices that are relevant to ten assigned attention objects distributed in six images. 2) Test at least 10 images freely to examine the synchronization results across devices. After the

Table 1. Average search time and recall.

Approach	Search Time(Second)	Recall(%)
Conventional	132.7	75.9
Synchronized	46.3	100

Table 2. Questionnaire results. (Notice, “it” refers to synchronized approach proposed in the paper. 1=strongly disagree, 5=strongly agree.)

Question	Rating
It greatly reduces the interactions across various devices when browsing images on them.	4.68
It should be an essential functionality in any device.	4.46
By using it, I find similar images across devices easier.	4.22
I would like to use it when I stay with my families or friends to share images together.	4.70
I feel satisfactory with the synchronization results it gives.	3.89
It improves the browsing experience across devices.	4.57

test is finished, each of the testers was asked to finish a questionnaire to rank the effectiveness of our approach to facilitate image browsing across devices.

Task 1 Observation. In this task, the subjects were divided into two groups, and each group consisted of four members with two females and two males. One group was assigned to use conventional browsing, and the other was left to use our synchronized way. Here, we adopt the average time reduction and recall as the measures to evaluate the performance. Recall is to represent the ratio of the number of correctly found images to the total number of target images. Notice the reason that we choose recall is to manifest that users commonly failed to find all of their wanted images for the inconvenience of interactions in small devices. The test results are listed in Table 1. As can be seen, the search time has been greatly reduced by about 64.6% in our approach, and the recall has also been improved to 100%. Obviously, testers can use synchronized image browsing to facilitate the search task prior to examining them one by one on the two devices. Furthermore, through observation of the search results by users, it’s found that the testers failed to seek all matched images, hence resulting in a lower recall.

Task 2 Observation. The subjects were first asked to select at least 10 images freely to examine the synchronization results. After such a task is finished, they were then asked to finish a questionnaire. The questionnaire asked for general information about our approach. The items are listed in Table 2. These questions were answered on a Likert scale, where 1=strongly disagree and 5=strongly agree. The table shows that users were overwhelmingly positive about our approach, claiming that such a service should be an essential functionality for any device. Note also that the rating of user satisfaction from the synchronization results is below 4, which remains to be improved in our future work. In general, users commonly thought ours is effective to improve the browsing experience.

6 Conclusions

In this paper, we proposed a novel synchronized approach to facilitate image browsing and searching across devices. The details of the synchronization across various devices are offered, including the system framework and the algorithm. With the satisfactory results from our use study, we are planning to extend our work to other media such as video.

References

1. L.Q. Chen, X. Xie, X. Fan, etc: A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal*, Vol. 9, No. 4, 2003, pp. 353-364.
2. F. Jing, M.J. Li, H.J. Zhang and B. Zhang: An effective region-based image retrieval framework. *ACM Multimedia 2002 Conference*, Dec. 2002, Nice, French.
3. H. Liu, X. Xie, W.Y. Ma and H.J. Zhang: Automatic browsing of large pictures on mobile devices. *ACM Multimedia 2003 Conference*, Nov. 2003, Berkeley, California, USA.
4. W.Y. Ma, I. Bedner, G. Chang, etc: A framework for adaptive content delivery in heterogeneous network environments. *MMCN 2000*, Jan. 2000, San Jose, USA.

An Intelligent Handoff Protocol for Adaptive Multimedia Streaming Service in Mobile Computing Environment

Jang-Woon Baek and Dae-Wha Seo

Electrical Engineering and Computer Science, Kyungpook National University,
1370 Sankyuk-Dong, Buk-gu, Dae-gu, Korea
kutc@palgong.knu.ac.kr, dwseo@ee.knu.ac.kr

Abstract. In this paper, we propose an intelligent hand-off protocol for adaptive multimedia stream service in mobile computing environment. In this protocol, the neighbor base stations receive media packets from a server in advance and transcode it to be suitable for mobile host. So, when a mobile host enters the neighbor cell, the base station immediately forwards the transcoded data packet to the mobile host. This paper describes the intelligent handoff protocol and the media data transcoding mechanism. We evaluate our approach with simulation results.

1 Introduction

Today, wireless internet users are growing dramatically and the demand of multimedia services are increasing. In mobile computing environment, the constraints of wireless links and the user mobility make it difficult to provide multimedia streaming services.

The wireless link has a low network bandwidth, high data loss rates, and frequent disconnection. The wireless network also has handoff problems which are occurred by the user mobility. When the mobile host enters a new cell, it needs a setup time for the new data path. If the handoff latency is long, it is difficult to provide seamless multimedia stream service because of the transmission delay and packet loss.

In the wireless network, the resources of wireless link are dramatically changed. It needs the adjustment of multimedia service quality. The media data are transcoded to meet the new service quality for the mobile host.

In this paper, we propose an intelligent hand-off protocol to support seamless multimedia stream service with less handoff latency and no packet loss. In this protocol, base stations which are neighborhoods to the primary base station, receive media data packets from a server and transcode the media data for mobile host, before the mobile host enters its cell. When the mobile host enters the neighbor cell, the neighbor base station immediately forwards the transcoded media data packet to the mobile host.

In the remainder of this paper, we present describe the intelligent handoff protocol in more detail and it is organized as follows. Section 2 discusses backgrounds for hand-off and transcoding. Section 3 describes the intelligent handoff

protocol. Section 4 presents the simulation results. Finally conclusion remarks are given in Section 5.

2 Background

To provide seamless multimedia service in wireless network, we have to ensure the user mobility and support the adaptive multimedia stream service based on the network condition and the device capacity of a mobile host.

Mobile IP[1] has been used to support the user mobility in mobile computing environments. In general, Mobile IP protocol has two steps for the routing of packets to a mobile host. In the first step, the packet is transferred from a correspond node(CN) to a home agent. In the second step, the home agent encapsulates packets and tunnels them to a foreign network. The foreign agent decapsulates the tunneled packets and forwards to the mobile host. During handoff, route update is performed. The packets are dropped or stored in buffer while the route is set up. Dropped packets cause the packet loss. The packet transmission after the completion of route-update brings about additional delay. These packet loss and delay degrade the quality of multimedia service. Therefore the new handoff protocol is require in order to achieve low handoff delay and negligible data loss. Handoff protocol using multicasting is proposed to overcome the problems caused by handoff[6]. This protocol achieves the fast handoff and reduces packet loss and handoff latency.

Multimedia data has the attributes of large volume and real time service. Accordingly, multimedia stream services require the wide bandwidth, the large memory and the high processing power. But it is difficult to provide the seamless multimedia stream services due to the narrow bandwidth and poor MN's device capacity in wireless network. To solve these problems, an intermediary has been located between wired network and wireless network. The intermediary plays a role such as filtering media data [2], transcoding web image [3], optimizing protocol [4], caching the part of media file [5] But most of these methods only consider characteristics of a physical link without reflection of user mobility. Although user mobility is supported, we can't say that the media streams are suitable for MH in new cell, because the transferred data is transcoded to meet the old cell environment while a mobile host moves a new cell.

3 Intelligent Handoff Protocol Environment(iHOPE)

3.1 iHOPE Architecture

In this paper, we propose the intelligent hand-off protocol environments(iHOPE) that supports the seamless multimedia streaming services ensuring user mobility. Figure 1 presents the iHOPE.

In the iHOPE, a mobile host (MH) is serviced from servers while the MH moves among the cells managed by base station(BS). If each BS senses that a new MH enters its own cell, the BS begins to receive media data

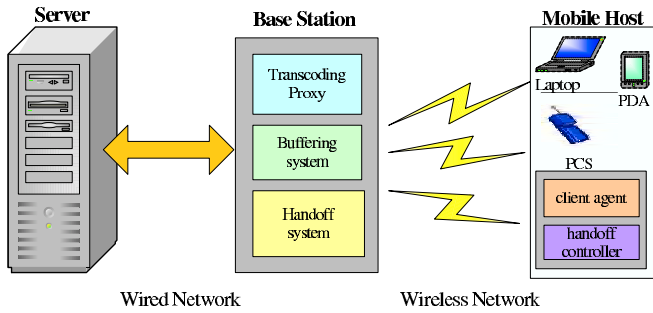


Fig. 1. Intelligent Handoff Protocol Environments

packets from the server and transcode the media data. The BS stores the transcoded data in buffer. When a BS performs the transcoding of media data, the BS considers the network bandwidth and the user profile including the device capacity and user preference of MH. So, if hand-off is completed, BS immediately sends the buffered data. Therefore, iHOPF ensures the user mobility and minimize transmission delay and packet losses during the hand-off. For the sake of these advantages, BS has a handoff module, a transcoding proxy and a buffering module. The MH has a client agent module and a handoff control module. The function of each module is described following.

- Buffering Module: To store data that are modified by transcoding proxy,
- Transcoding Proxy: to transcode the media data according to the user profile and the bandwidth of network,
- Handoff Module: to support user mobility,
- Handoff control Module: to decide the route and control BS during handoff,
- Client module: to provide the user profile and control the transcoding proxy according to the change of host's environment.

3.2 iHOPE Handoff Protocol

iHOPE network stack for intelligent handoff is depicted in figure 2.

When a MH resides in foreign network, a home agent encapsulates multimedia packets with multicast IP address associated with mobile host(MH). BSs that associated with the MH mean the BS forwarding packets to MH or the neighborhood BS. These BSs should join the same multicast group. Encapsulated packets are multicasted into foreign network. A Decapsulation Module extracts multimedia data from packets that are transmitted from HA, and gives multimedia data to the transcoding proxy. A Transcoding proxy transcodes the media data based on network bandwidth and the MH's profile. A forwarding BS transmit the transcoded packet to the MH, while neighbor BSs store the transcoded packets. BS's beacon module keeps track of the roaming MH. Each BS periodically broadcasts a beacon message to mobile hosts. BS's buffering module

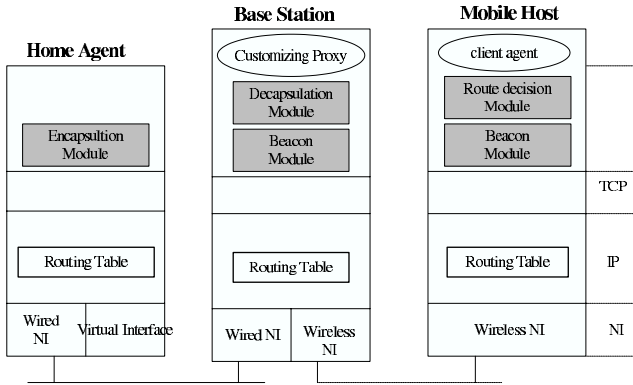


Fig. 2. iHOPE Network Stack

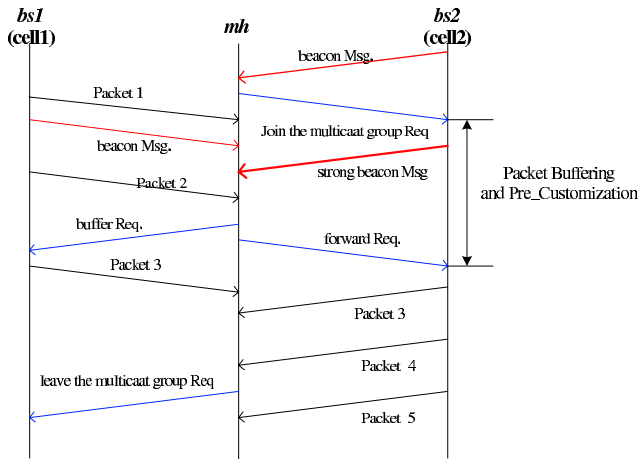


Fig. 3. iHOPE Handoff Flow

manages packets that are received from HA or modified by transcoding proxy. The MH’s beacon module informs the BS about the MH’s current location and communication states and the route decision module decides a forwarding BS and a buffering BS based on a BS’s received signal strength and communication states.

Figure 3 depicts a handoff protocol between BSs and a MH when a mh moves from the cell1 to the cell2.

The *mh* periodically receives beacon messages from BSs(*bs1*, *bs2*). As the *mh* approaches to the cell 2, the *mh* gets more strong signals than before. If the *mh* receives the stronger signal than the threshold of signal strength, the *mh* name the *bs2* for buffering base station. At same time the *mh* sends the message that requests the *bs2* to join the multicast group. Then the *bs2* joins

the *mh*'s multicast group and receives packets from the home agent. The *bs2* decapsulates the received media packets, and transcodes media data through transcoding proxy, and then stores the transcoded media data packets in the buffer. If the *bs2*'s signal strength is stronger than the *bs1*'s signal strength, the *mh* decides the forwarding BS with the *bs2* and requests the packet forwarding. The *bs2* immediately sends the transcoded packet which is stored in buffer when *bs1* enters the cell2. At this time, *bs1* receives the buffer request and becomes the buffering BS. If *mh* goes away from *bs2* and receive weaker signal than the lower threshold, *mh* sends a message that asks *bs1* to leave the multicast group. *bs1* no longer receives the message.

3.3 Transcoding Proxy

BS's Transcoding Proxy(TP) transcodes media data based on network bandwidth and user profile(user preference, MN's device capacity). Figure 4 shows the internal architecture of TP. TP consists of policy manager, transcoding module, network monitor, user profile database, front door.

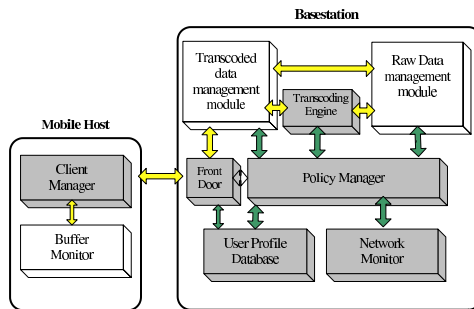


Fig. 4. Internal Structure of Transcoding Proxy

Policy manager(PM) makes decision of transcoding policy on the basis of MH's user profile and network monitoring information. Transcoding policy means the set of transcoding parameters such as encoding bit rate, frame size, frame format, quantization table.

Transcoding engine(TE) performs the transcoding of media data according to transcoding policy. TE decodes the media data, filters it based on policy and encodes.

The network monitor(NM) observes the change of resource between MH and BS. NM collects the latency and the available bandwidth between MH and BS. NM gives the network information to the policy manager through query based interface. Figure 6 shows the internal block diagram of network monitor.

User Profile is stored in database. User profile includes a user information for authentication, the device capacity and user preference of mobile host.

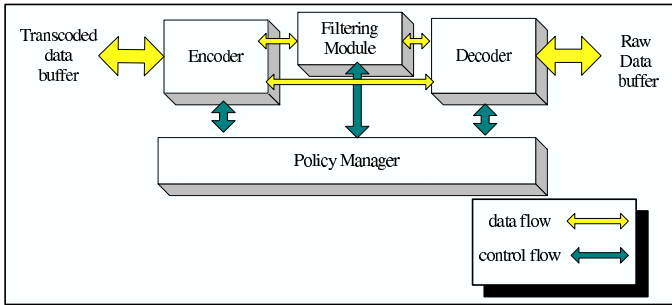


Fig. 5. Internal Structure of Transcoding Proxy

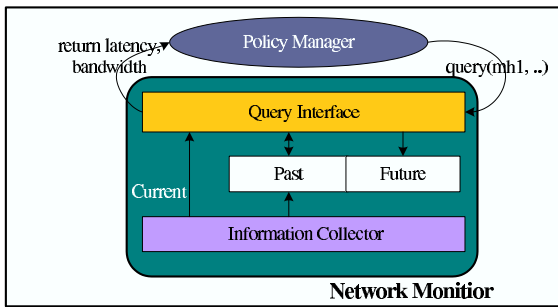


Fig. 6. Internal Structure of Transcoding Proxy

Frontdoor(FD) is responsible for MH’s authentication using user profile in the database and intermediation between the server and the MH. When MH wants to be served the multimedia stream, TP’s FD certifies the user ID and the user password. If MH is authenticated, the FD asks a server to provide the media stream. And then FD receives the media stream from the server. The transcoding proxy retrieves the stream and modifies it and sends it to FD, which passes it to the user or stores it in the transfer buffer.

3.4 Buffering Module

BS’s buffering module manages the media data which are received from server and are transcoded by transcoding proxy. Buffering module consists of RDMM(raw data management module) and TDMM(transcoded data management module).

RDMM extracts the media data from the packet sent from server and stores the data in the raw data buffer. TDMM stores the transcoded data in the transfer buffer or cache transcoded data in the disk. When MH enters the BS’s cell, the data in transfer buffer are immediately sent to MH. When a mobile host requests multimedia data, a transcoding proxy check the media cache. If cache

hit is occurred, the proxy directly sends the transcoded media data to the mobile host. Otherwise, the proxy ask server to send multimedia data.

4 Simulation

We used the *ns*(network simulator)[5] to evaluate the iHOPE. We design the simulation model like Fig. 7. We design and implement the iHOPE handoff protocol. And we create new multimedia application to show the transcoding proxy’s role. This application adjusts the transmission rate according to the network environments. We evaluate two cases of simulation. One is the thing which applies the iHOPE handoff protocol, another is the thing which uses general handoff protocol. We used the simulation to compute the three performance metrics: the *handoff latency*, the *number of packet loss* and the *jitter*.

We induced the handoff latency and the number of packet loss from the ns-trace file. That is very easy; the handoff latency is the difference of the last packet arrived time in the old base station and the first packet arrived time in the new base station; the number of packet loss is the difference of the last packet sequence number in the old base station and the first packet sequence number in the new base station.

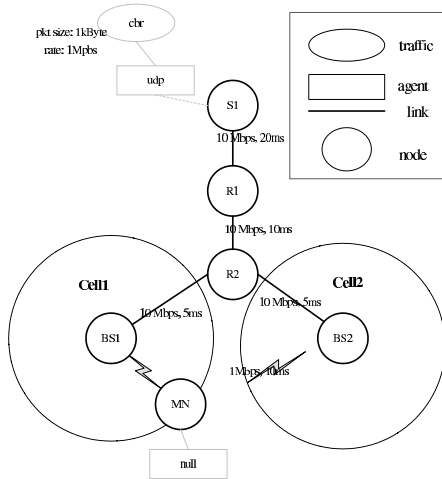


Fig. 7. Simulation Model

Table 1. The Result of Network Simulation

Handoff Protocol	Handoff Latency	Number of Packet Loss	Jitter
General Handoff Protocol	10s	2400	0.017Mbps
iHOPE Handoff Protocol	1.1s	199	0.01Mbps

Table 1 shows the result of simulation. We measured the handoff latency, number of packets lost and the average jitter. iHOPE handoff Protocol has lower handoff latency and jitter than the general handoff protocol. Also there isn't the packet loss in iHOPE.

5 Conclusion and Future Work

In this paper, we proposed iHOPE to reduce handoff latency and packet loss, and to overcome constraint of wireless link and MH's device capacity. In iHOPE each BS buffers the packet for MH and transcodes the media data before mobile host enter the cell. As a result, iHOPE is able to support seamless multimedia stream service for MH with less handoff latency and packet loss. The result of simulation shows the excellence of iHOPE in term of packet loss and handoff latency.

In the future, we study the media caching to reduce the response time for user. And we study transfer policy that considers the priority of application.

References

1. Charles E. Perkins: "Mobile IP", IEEE Communication Magazine, (1997)
2. Bruce Zenel: "A general purpose proxy filtering mechanism applied to the mobile environment", IEEE Wireless Networks 5 pp.391-409, 1999
3. Srinivasan Sesha et al.: "Handoffs in Cellular Wireless Network: The Daedalus Implementation and Experience", Kluwer International Journal on Wireless Personal Communication, 1997.1
4. Armondo Fox et al.: "Adapting to network and client variability via On Demand Dynamic Distillation", Proceeding. Seventh International. Conference. on Architecture. 1996.10
5. M. Liljeberg et al.: "Optimizing World-Wide Web for Weakly Connected Mobile Workstations: An indirect approach", Proceedings of SDNE'95, pp.132-139, 1995.6
6. H. Fabmi et al.: "Proxy Server for Scalable Interactive Video Support", IEEE Computer, Vol.34, pp.54-60, 2001.9
7. the network simulator: ns-2, <http://www.isi.edu/nsnam/ns/>, (current October 2003)

Platform Architecture for Seamless MMS Service over WLAN and CDMA2000 Networks

Su-Yong Kim¹, Yong-Bum Cho¹, and Sung-Joon Cho²

¹ Department of Information and Telecommunication Engineering,
Graduate School of Hankuk Aviation University, Korea
{sykim, athome}@mail.hankong.ac.kr

² School of Electronics, Telecommunication, and Computer Engineering,
Hankuk Aviation University, Korea
sjcho@mail.hankong.ac.kr

Abstract. Multimedia Messaging Service (MMS) has been seen as the key application in its entry into 3G services. Recently, the combination of WLAN and 3G wireless technologies will make MMS service more ubiquitous, bringing benefits to both service providers and their customers. To realize ubiquitous MMS service over WLAN and CDMA2000 networks, we design and implement new platform architecture by reusing the existing standards and network elements at the same time. We employ loose coupling approach and Mobile IP approach to propose new platform architecture, interfacing MMS Center (MMSC) with many existing components such as SMSC, PPG, HLR, HA, AAA, PDSN, WLAN Gateway, and so on. Based on our proposed platform architecture, we also present seamless MMS delivery scenarios that can't be possible within the current MMS reference architecture. This paper will make a contribution for service providers to provide their customers with seamless MMS service over WLAN and CDMA2000 networks.

1 Introduction

The combination of WLAN and 3G wireless technologies offers the possibility of achieving anywhere and anytime Internet access, bringing benefits to both service providers and their customers. Given the complementary characteristics of WLAN (faster short-distance access) and CDMA2000 (slower long-distance access), it is compelling to combine them to provide ubiquitous wireless access. It will allow CDMA2000 service providers to economically offload data traffic from wide-area wireless spectrum to WLAN indoor locations, hotspots, and other areas with high user density. For WLAN service providers, the integration will bring them a larger user base from CDMA2000 network, without having to win them through per-customer service contractors [1,2].

Multimedia Messaging Service (MMS) provides the means for delivering multimedia messages between two mobile stations in store-and-forward fashion. Furthermore, MMS provides mobile users with the possibility to exchange multimedia messages with the Internet E-mail users. The 3GPP [3], 3GPP2 [4], WAP

Table 1. Media formats and types for MMS

Media Formats	Media Types in 3GPP	Media Types in 3GPP2
Text	M:US-ASCII, ISO-8859-1, UTF-8, Shift_JIS, etc	M:US-ASCII, ISO-8859-1, UTF-8, Shift_JIS, GSM 7-bit default alphabet, etc
Speech	M:AMR, AMR-WB	M:3GPP2 13K, AMR
Audio	O:MPEG-4AAC LC,MPEG4-AAC LTP	O:MPEG-4AAC LC,MPEG-4AAC LTP
Synthetic Audio	O:SP-MIDI	O:SP-MIDI
Still Image	M:JPEG, JFIF	M:JPEG, JFIF
Bitmap Graphics	O:GIF, PNG	O:GIF, PNG
Video	M:H.263 Profile 0 Level 10 O:H.263 Profile 3 Level 10, MPEG-4 SP Level 0	M:H.263 Profile 0 Level 10, MPEG-4 SP Level 0 O:H.263 Profile 3 Level 10
Vector Graphics	M:SVG-Tiny O:SVG-Basic	M:SVG-Tiny O:SVG-Basic
File Format for Dynamic Media	M:MP4	M:.3g2
Media Synchronization and Presentation Format	M:SMIL O:XHTML Mobile Profile	M:SMIL

M: Mandatory, O: Optional

Forum [5], and OMA [6] standardize MMS. For the sake of interoperability and alignment, they also specify media formats and types as listed in Table 1.

A network element called MMS Center (MMSC) implements the store-and-forward functionality in CDMA2000 network. In CDMA2000, with a Mobile Station ISDN Number (MSISDN), MMSC notifies the recipient Mobile Station (MS) of the new MMS message through Short Message Service (SMS) as the bearer protocol via SMS Center (SMSC). The notification contains the message size, the message reference (URL), the message subject, the originator address, etc. On the other hand, in WLAN, there is no specific method to send the notification about the incoming MMS message to the recipient MS which is identified with an IP address, not an MSISDN. This means that MMSC should inform the recipient MS of the new MMS message with an IP address in WLAN. As this functionality is not supported by the current MMSC, MMS service is not provided when the recipient MS is in WLAN, in spite of the integration of WLAN and CDMA2000.

In this paper, we design new platform architecture for supporting seamless MMS service over WLAN and CDMA2000 networks by reusing the existing standards and network elements at the same time. We employ loose coupling approach and Mobile IP approach to propose new platform architecture, interfacing MMSC with many existing components such as SMSC, Push Proxy Gateway (PPG), Home Location Register (HLR), Home Agent (HA), Authentication Authorization Accounting (AAA), Packet Data Service Node (PDSN), WLAN Gateway, and so on. Based on our proposed platform architecture, we also present seamless MMS delivery scenarios that can't be possible within the current MMS reference architecture.

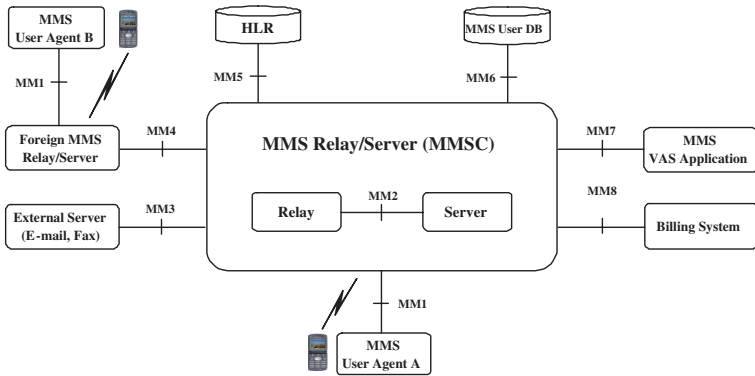


Fig. 1. General view of MMS reference architecture.

2 MMS Reference Architecture

As shown in Fig. 1, MMS reference architecture contains several key elements that interwork with one another to provide MMS service. The key elements defined by the 3GPP, 3GPP2, WAP Forum, and OMA are described as followings.

MMS User Agent: MMS User Agent (UA) is responsible for composing and rendering multimedia messages. Message rendering is performed by using the appropriate content rendering service. MMS UA also sends and receives multimedia messages with the appropriate network protocols.

MMS Relay/Server (MMSC): MMS Relay, which is the key element in MMS Environment (MMSE), takes the responsibility for transferring multimedia messages across different messaging systems, transcoding multimedia message format, interworking with other platforms, and enabling access to various servers residing in other networks. MMS Server stores and handles the incoming/outgoing multimedia messages.

MMS User DB: MMS User Database (DB) may consist of lots of different data, such as user profile database, subscription database, and HLR information for mobility management.

MMS VAS Application: MMS Value Added Service (VAS) Application could be included in or connected to MMSE. The MMS VAS Application offers a value added service to MMS users and then may be able to generate charging data.

External Servers: Several external servers, such as E-mail server, Fax server, and SMSC, can be included in MMS reference architecture. Convergence functionality between external servers and MMS UA is provided by MMSC, which enables the integration of different server types across different networks.

3 Proposed MMS Platform Architecture

MMS delivery between two mobile stations can be divided in two parts. In the first part, referred to Mobile Originated (MO) message delivery, the originator MS submits the message to MMSC. In the second part, after completely receiving the message from the originator MS, MMSC delivers the message to the recipient MS. This part is referred to Mobile Terminated (MT) message delivery. An MS is identified with an MSISDN in CDMA2000 and with an IP address in WLAN. Currently in CDMA2000, with an MSISDN, MMSC informs the recipient MS of the new MMS message through SMS via SMSC. In WLAN, however, there is no specific method to send the notification about the incoming MMS message to the recipient MS which is identified with an IP address, not an MSISDN. Therefore, MMS service can't be supported to the recipient MS in WLAN, in spite of the integration of WLAN and CDMA2000.

As illustrated in Fig. 2, we employ loose coupling [7,8] approach and Mobile IP [9] approach to propose new platform architecture for supporting seamless MMS service over WLAN and CDMA2000 networks. We interface MMSC with many existing network elements such as SMSC, PPG, HLR, HA, AAA, PDSN, WLAN Gateway, and so on. When the recipient MS is in WLAN, MMSC needs to find out an IP address of the recipient MS with the MSISDN provided by the originator MS. First, MMSC asks for an International Mobile Subscriber Identity (IMSI) to HLR with the MSISDN. And then, MMSC requests the IP address to HA with the IMSI derived from HLR. With the IP address obtained

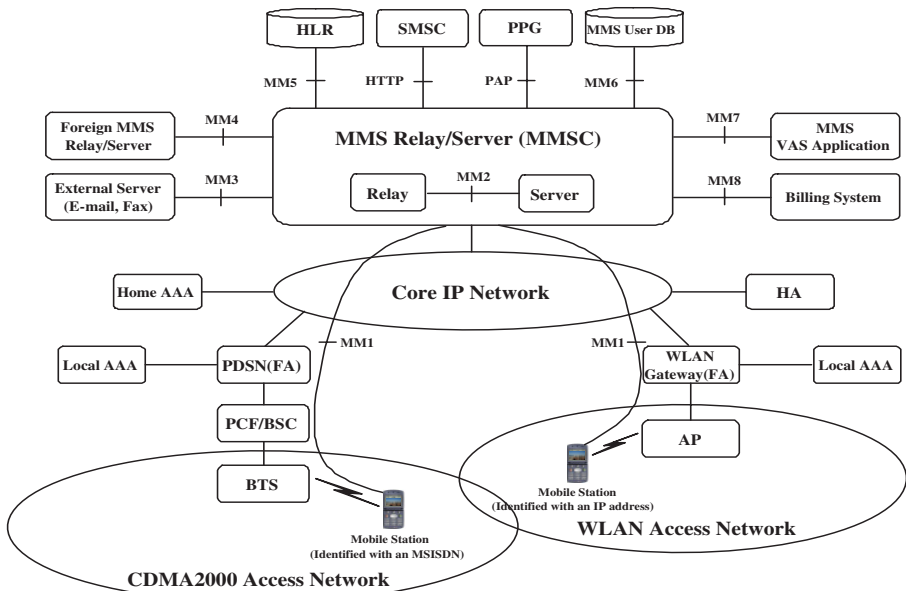


Fig. 2. Platform architecture for seamless MMS service over WLAN and CDMA2000.

from HA, MMSC sends the notification about the new MMS message to the recipient MS through HTTP via PPG. The new functions and elements adapted in our proposed platform architecture are described as followings.

Push Initiator: MMSC works as a Push Initiator (PI) which originates push content and submits it in the form of a push request using Push Access Protocol (PAP) to PPG for delivery to the recipient MS. Each submission has a unique identifier. MMSC can request the result of a push submission.

PPG: PPG is responsible for delivering the push content to the recipient MS. In doing so, it potentially may be needed to translate the recipient address provided by MMSC into an address form understood by the mobile network, to modify the push content taking into consideration of the recipient MS's capabilities, to store the content when the recipient MS is unavailable to receive the notification, etc. In addition to push delivery, PPG may notify MMSC about the result of a push submission.

PAP: PAP is based on standard Internet protocols. XML is used to express the delivery instructions and the push content can be any MIME media type. These standards can make WAP Push more flexible and extensible [10].

4 Seamless MMS Delivery Scenarios

To implement our new platform architecture, we have considered seamless MMS delivery scenarios illustrated in Fig. 3 and Fig. 4 that can't be possible within the current MMS reference architecture.

1. To deliver a new MMS message, an originator MS submits MM1_submit.req message to MMSC using HTTP Post method in WLAN and CDMA2000.
2. MMSC sends MM1_submit.res message back to the originator MS indicating whether the new message is successfully accepted or not.
3. As the notification mechanism is dependent on the location of the recipient MS, MMSC needs to check which network the recipient MS is connected to before notifying the recipient MS of the incoming MMS message. In CDMA2000 like Fig. 3, MM1_notification.req message can be sent to the recipient MS with the MSISDN through SMS via SMSC. On the other hand, in WLAN, MMSC needs to find out an IP address of the recipient MS from HA with the MSISDN provided by the originator MS. In WLAN like Fig.4, MM1_notification.req message can be delivered to the recipient MS through HTTP via PPG.
4. The recipient MS responds with MM1_notification.res message using HTTP Post method in WLAN and CDMA2000.
5. To retrieve the new MMS message, the recipient MS transmits MM1_retrieve.req message to MMSC using HTTP Get method with the URL provided by MM1_notification.req message in WLAN and CDMA2000.

6. According to the terminal capabilities described in MM1.retrieve.req message, MMSC customizes the multimedia contents of the new MMS message. After encapsulating the new MMS message into MM1.retrieve.res message, MMSC sends it to the recipient MS using HTTP Post method in WLAN and CDMA2000.

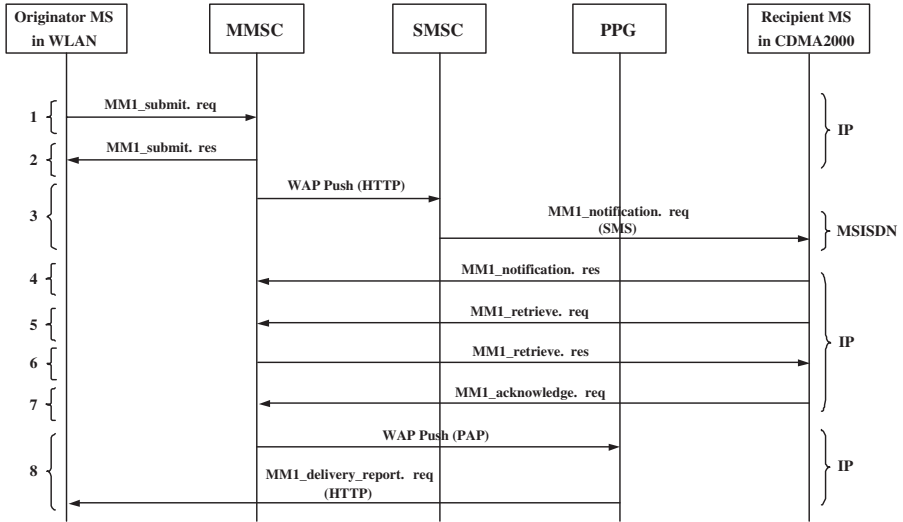


Fig. 3. Seamless scenario for MMS delivery from WLAN to CDMA2000.

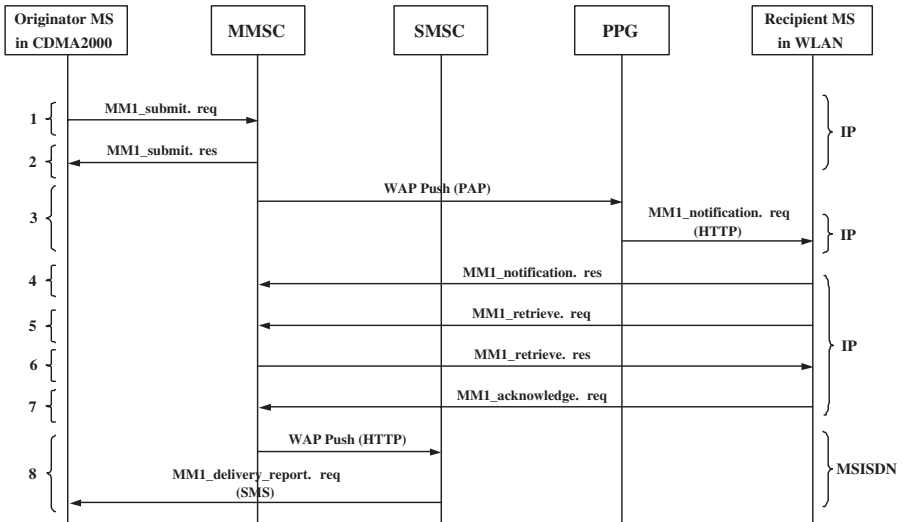


Fig. 4. Seamless scenario for MMS delivery from CDMA2000 to WLAN.

7. On receiving MM1_retrieve.res message, the recipient MS acknowledges it by sending MM1_acknowledge.req back to MMSC in WLAN and CDMA2000.
8. Before forwarding the delivery report to the originator MS, MMSC also needs to check the location of the originator MS. If the originator MS is in CDMA2000 like Fig. 4, MM1_delivery_report.req message can be sent to the originator MS with the MSISDN through SMS via SMSC. However, when the originator MS is connected to WLAN, MMSC first has to obtain an IP address of the originator MS from HA. And then, as illustrated in Fig. 3, MM1_delivery_report.req message can be submitted to the originator MS with the IP address through HTTP via PPG.

If WLAN doesn't support MMS service, MMSC has to delay the notification about the incoming MMS message or the delivery report until the recipient MS or the originator MS moves into CDMA2000 from WLAN.

5 Conclusions

In this paper, we have designed the new platform architecture for supporting seamless MMS service over WLAN and CDMA2000 networks while considering the possibility of reusing the existing standards and network elements at the same time. We have employed loose coupling approach and Mobile IP approach to propose new platform architecture, interfacing MMSC with many existing components such as SMSC, PPG, HLR, HA, AAA, PDSN, WLAN Gateway, and so on. Based on our proposed platform architecture, we have also presented seamless MMS delivery scenarios that can't be possible within the current MMS reference architecture. This paper will make a contribution for service providers to support the seamless MMS service over WLAN and CDMA2000 networks. For secure MMS service, we need more investigations in the area of common authentication and mobility management.

Acknowledgment. This research was supported by IRC (Internet Information Retrieval Research Center) in Hankuk Aviation University. IRC is a Kyounggi-Province Regional Research Center designated by Korea Science and Engineering Foundation and Ministry of Science & Technology.

References

1. M. M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, S. Miller, and L. Salgarelli, "Design and implementation of a WLAN/CDMA2000 interworking architecture," *IEEE Communications Magazine*, vol. 44, pp. 90-100, November 2003.
2. M. Cappiello, A. Floris, and L. Veltri, "Mobility amongst heterogeneous networks with AAA support," *ICC 2002, IEEE International Conference on*, pp. 2064-2069, May 2002.
3. 3GPP TS 23.140, *Multimedia Messaging Service (MMS), Functional Description; Stage 2*, 3GPP, September 2002.

4. 3GPP2 X.S0016.200, MMS Stage2, Functional Description, Revision: 0, 3GPP2, April 2003.
5. WAP-205-MMSArchOverview-20010425-a, Multimedia Messaging Service Architecture Overview Specification, WAP Forum, April 2001.
6. OMA-MMS-ARCH-v1_2-20031217-c, MMS Architecture Overview, OMA, December 2003.
7. K. Ahmavaara, H. Haverinen, and R. Pichna, "Interworking architecture between 3GPP and WLAN systems," *IEEE Communications Magazine*, vol. 44, pp. 74-81, November 2003.
8. A. K. Salkintzis, "Interworking between WLANs and third-generation cellular data networks," *VTC 2003-Spring*, vol. 3, pp. 1802-1806, April 2003.
9. C. Perkins, "IP Mobility Support," *IETF RFC 2002*, October 1996.
10. WAP-250-PushArchOverview-20010723-p, WAP Push Architectural Overview Specification, WAP Forum, July 2001.

Make Stable QoS in Wireless Multimedia Ad Hoc Network with Transmission Diversity

Chao Zhang¹, Mingmei Li²,
Xiaokang Lin¹, Shigeki Yamada², and Mitsutoshi Hatori²

¹ Dept. of Elec. Engr. of Tsinghua University, Beijing, 100084, P.R.China
zhangchao2000@mails.tsinghua.edu.cn,

² National Institute of Informatics (NII), Chiyada, Tokyo, 101-8430, Japan
shigeki@nii.ac.jp

Abstract. Providing stable Quality of Service (QoS) of multimedia transmission in wireless mobile Ad Hoc network is still a challenging problem. In this paper, we present a novel transmission diversity scheme which uses multiple transmitting nodes and space-time block codes in physical layer. Multi-path route is generated in network layer. This scheme has been proved to be an efficient way to offer stable QoS in wireless mobile environment. Specifically, the network equipped with the new scheme can provide the same Symbol Error Rate (SER) in physical layer, whereas the total energy consumption of transmission is decreased by diversity transmission.

1 Introduction

Multimedia service in wireless Ad Hoc networks shows promising applications in next generation wireless communications, ranging from civil projects, monitoring to military etc [1]. Due to the mobility and the instability of wireless channel, providing stable Quality of Service (QoS) in Ad Hoc network is difficult. It becomes an apparent bottleneck to make multimedia transmission, which is sensitive to QoS, feasible in such kind of networks. On the other hand, Energy issue is also a challenging problem. Usually, the energy of the nodes in Ad Hoc network is limited. With the energy exhausting, high Symbol Error Rate (SER) occurs and QoS almost can not be maintained.

Therefore, the question can be concluded as we should find a way to offer the same QoS when the energy of the nodes are going down. Specially, in physical layer, we should find a new scheme to maintain SER with less energy. In our paper, we only take into account the transmission energy which is the major energy consumption of the nodes in Ad Hoc network.

In this paper, the above question is solved by multi-path route based on transmission diversity in the physical layer. The performance is considerably good according to the simulation. The rest of the paper is organized as follows: In Section 2, we show the diversity transmission in physical layer. Then, we present the algorithm to generate multi-path route in Section 3. In Section 4, we analyze the performance of the algorithm and multi-path rout by simulation. Finally, we briefly summarize this paper and give a conclusion in Section 5.

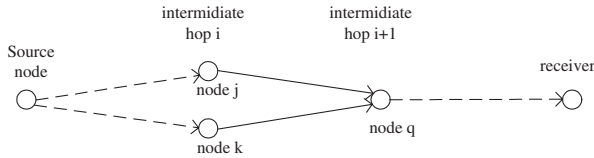


Fig. 1. Multi-hop wireless sensor network

2 Diversity Transmission in Ad Hoc Network

2.1 Diversity in Physical Layer

In wireless Ad Hoc networks, there exists high possibility that each transmission by a node results in simultaneous reception by multiple other nodes. Thus, it is possible for the data packet to be re-transmitted simultaneous by multiple nodes. This diversity scheme is proved to offer energy efficiency in transmission compared with the traditional schemes are “single transmission” where only one node is selected to perform transmission per hop per routing path [2]

In this paper, a simple two-node diversity scheme is involved. The fundamental principle is originated from Space-Time Block Code (STBC) in sensor networks [3]. It can be illustrated in Fig. 1, where a source node needs to transmit data packets to the remote receiver through a multi-hop wireless network.

In the intermediate hop i , a data packet from the source node is received by two nodes, e.g. node j and node k . In traditional single transmission scheme, only one node is chosen to retransmit the data packet to hop $i+1$. However, in this scheme, two of the nodes in hop i are chosen to perform diversity transmission with STBC [4].

Without loss of generality, assume nodes 1 and 2 have both received data packets from the previous hop as

$$P_1 : \{s_1(0), \dots, s_1(N - 1)\}, P_2 : \{s_2(0), \dots, s_2(N - 1)\} \tag{1}$$

where $s_i(n), i = 1, 2$ are uniformly distributed symbols. Then, the packets (PN_1 from node 1 and PN_2 from the node 2) will be transmitted to next hop (node q in Fig.1).

$$\begin{aligned} PN_1 &: \{s_1(0), \dots, s_1(N - 1), -s_2^*(N - 1), \dots, -s_2^*(0)\} \\ PN_2 &: \{s_2(0), \dots, s_2(N - 1), s_1^*(N - 1), \dots, s_1^*(0)\} \end{aligned} \tag{2}$$

where “*” denotes complex conjugate.

Assume channels from nodes 1 and 2 to a receiver in the next hop are Rayleigh flat-fading with coefficients α_1 and α_2 . Let the transmission delays be d_1 and d_2 respectively. In [3], a simple expression is given for the packet received in the hop $i + 1$ (node q) indicated by $\mathbf{x}(n)$.

$$\begin{aligned} \mathbf{x}(n) &= \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_2^* & -\alpha_1^* \end{bmatrix} \begin{bmatrix} s_1(n - d_1) \\ s_2(n - d_2) \end{bmatrix} + \begin{bmatrix} v_1(n) \\ v_2^*(N - n - 1 + d_1 + d_2) \end{bmatrix} & (3) \\ &= \mathbf{H}\mathbf{s}(n) + \mathbf{v}(n) & (4) \end{aligned}$$

where $v_1(n)$ and $v_2(n)$ are zero-mean additive white Gaussian noise (AWGN).

With the estimated d_1, d_2 and channels α_1, α_2 , we can estimate symbol vector $\mathbf{s}(n)$ by maximum likelihood detection as

$$\hat{\mathbf{s}}(n) = (|\alpha_1|^2 + |\alpha_2|^2)^{-1} \mathbf{H}^H \mathbf{x}(n) \tag{5}$$

where $(.)^H$ denotes conjugate transpose. This standard STBC codes transmission scheme achieves full diversity and full rate [4].

Simulation result is shown in [3]. The diversity transmission scheme has lower Signal to Noise Ratio (SNR) with the same SER. It shows that the diversity scheme required 5dB less SNR to achieve SER 1%, thus the transmission can have the same QoS performance even the energy decreased three times.

2.2 Multi-path Route in Topology

In this paper, we assume all nodes are located in the two dimensional plane. The power function or the transmit power for node u to reach node v is denoted by $t \cdot d^\alpha$, where d is the distance between u and v , t is the threshold which is a function of the signal-to-noise ratio at v , and α is a constant that is related to path loss. In a typical application, α is from two to four [5]. We use $f(d)$ to indicate the power function assigned to a node with transmission distance d .

If node u and v set up a single path transmission link, the transmission power of u is denoted as $f_1(u, v)$. If node u and v set up a link in diversity transmission, the transmission power of u is denoted as $f_2(u, v)$. From Subsection 2.1, we can see at least

$$f_1(u, v) > 2f_2(u, v) \tag{6}$$

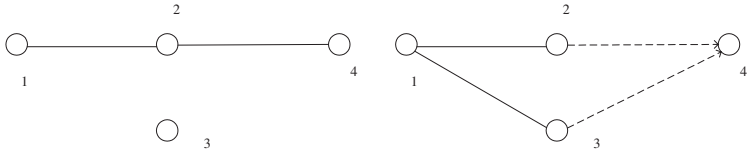
When BER=1%, the simulation result even shows

$$f_1(u, v) > 6f_2(u, v) \tag{7}$$

In general, if we employ the transmission diversity in physical layer, the routing topology becomes into multi-path too. For example, in Fig.2(a), node 1 can send data packet to node 4 via node 2 if it is the single path transmission. The Minimum Energy Topology (MET) [6] is highlighted. Whereas, if there is the condition that

$$f_1(1, 2) + f_1(2, 4) > \max\{f_1(1, 2), f_1(1, 3)\} + f_2(2, 4) + f_2(3, 4) \tag{8}$$

In order to get lower energy, we should change the routing topology into diversity transmission shown in Fig.2(b). where solid lines indicate the links with transmission power of $f_1(d)$, the dotted lines indicate the links with transmission power of $f_2(d)$. Node 1 transmit diversity data to node 2 and node 3 with the



(a) Single transmission topology (b) Diversity transmission topology

Fig. 2. Routing topology comparison of single and diversity transmission

transmission power of $max\{f_1(\{1, 2\}), f_2(\{1, 3\})\}$. Then, node 2 and node 3 relay the data to sensor 4 by physical diversity narrated in above subsection.

Above description indicates we can maintain the QoS, actually SER in physical layer, in a same level with less energy consumption in transmission.

3 Algorithm to Construct Multi-path Route with Diversity Transmission

3.1 Algorithm

1. Find the network topology with strong connection, i.e. bidirectional link between the nodes. We denote the topology as $G = (V_G, E_G)$, where V this the set of nodes and E is the set of the links (edges) between the nodes.
2. Produce Minimum Energy Topology (MET). Usually, we find the Minimum Spanning Tree (MST) [2]. The node on the tree can reach the farthest neighbor with minimum energy along the route. We denote the topology as $T = (V_T, E_T)$. There are lots of algorithms to get MST from above network topology G, e.g. Kruskal’s algorithm or incremental power greedy heuristic algorithm [6].
3. We assume there is a monitoring node (indicated as monitor) in the wireless Ad Hoc network and all the nodes send data to it. We can easily find the minimum energy route from node to monitor on T. Without loss of generality, we assume it as (n_1, n_2, \dots, n_N) , where n_1 is the node send out original data packet, n_N is the monitor and $\{n_2, \dots, n_{N-1}\}$ is the relay nodes. Now, we can construct the energy efficient multi-path route based on this single transmission route.

Input: Route set R, Node set S and Power set P are initialized. The initial values are $R = \{(n_1, n_2)\}$, $S = \{n_1, n_2\}$ and $P = f_1(n_1, n_2)$

Output: Energy efficient multi-path route with transmission diversity R and the transmission power assigned to each node.

Heuristic algorithm:

- a) $i=1$
- b) if $N=2$ then stop. Otherwise, continue with (c)
- c) if \exists node j , (n_i, j) and $(j, n_{i+2}) \in G$ and

$$f_1(n_i, n_{i+1}) + f_1(n_{i+1}, n_{i+2}) > \min\{\max\{f_1(n_i, n_{i+1}), f_1(n_i, j)\} + f_2(n_{i+1}, n_{i+2}) + f_2(j, n_{i+2})\} \quad (9)$$

when $j = j_i$, min is available.

Then, we modify S,R,P as

$$S = S \cup \{j_i, n_{i+2}\} \quad (10)$$

$$R = R \cup \{(n_i, j_i), \langle n_{i+1}, n_{i+2} \rangle, \langle j_i, n_{i+2} \rangle\} \quad (11)$$

$$P = P + f_2(n_{i+1}, n_{i+2}) + f_2(j_i, n_{i+2}) + \max\{f_1(n_i, n_{i+1}), f_1(n_i, j_i)\} - f_1(n_i, n_{i+1}) \quad (12)$$

where (x, y) indicates the link with transmission power $f_1(x, y)$ and $\langle x, y \rangle$ indicates the link with transmission power $f_2(x, y)$.

Otherwise, we modify S,R,P as

$$S = S \cup \{n_{i+2}\} \quad (13)$$

$$R = R \cup \{(n_{i+1}, n_{i+2})\} \quad (14)$$

$$P = P + f_1(n_{i+1}, n_{i+2}) \quad (15)$$

- d) If n_{N-1} is already belongs to V ($n_{N-1} \in V$), then stop. Otherwise, $i=i+1$ and go to (c).

3.2 Example

To demonstrate our algorithm, we provide the following example.

Assume the SER limitation is 1% according to the requirement of QoS, $f_1(d) = t_1 d^\alpha$, $f_2(d) = t_2 d^\alpha$. According to Eq.7, $t_1 = 6t_2$ (for one link, not for the total). Then, $f_1(d) = 6f_2(d)$. at the same time, we assume $\alpha = 2$.

The network topology of the example is shown in Fig.3, where the number on the edge indicates the distance between the adjacent two nodes. The node 1 sends data packet to the node 9 which is actually the monitor node of the network.

First, we find the network topology with strong connection $G = (V_G, E_G)$. This topology is already illustrated in Fig.3 by nodes and edges.

Second, we search the Minimum Spanning Tree (MST) $T = (V_T, E_T)$ on the above topology. This tree is also highlighted out in Fig.3 by bold solid lines and all the nodes. Then it is easy to confirm the minimum energy route from the node 1 to node 9 is $R = \{(n_1, n_2), (n_2, n_4), (n_4, n_6), (n_6, n_8), (n_8, n_9)\}$.

Third, we start to search the multi-path transmission route based on above single transmission route by heuristic algorithm.

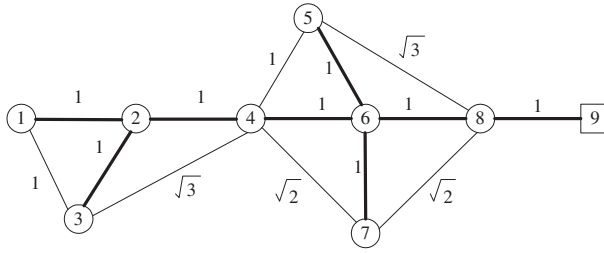


Fig. 3. Network topology of the example

The steps are described below:

1. $i=1, S = \{n_1, n_2\}, R = \{(n_1, n_2)\}$ and $P = f_1(1)$.

Since

$$f_1(n_1, n_2) + f_1(n_2, n_4) = 2f_1(1) = 12f_2(1) = 12t_2 \tag{16}$$

and

$$\begin{aligned} & \max\{f_1(n_1, n_2), f_1(n_1, n_3)\} + f_2(n_2, n_4) + f_2(n_3, n_4) \\ &= \max\{f_1(1), f_1(1)\} + f_2(1) + f_2(\sqrt{3}) \\ &= f_1(1) + f_2(1) + f_2(\sqrt{3}) \\ &= 10t_2 \end{aligned} \tag{17}$$

Then, $Eq.16 > Eq.17$. Therefore, we choose multi-path transmission route in this step.

2. $i=2, S = \{n_1, n_2, n_3, n_4\}, R = \{(n_1, n_2), (n_1, n_3), (n_2, n_4), (n_3, n_4)\}$ and $P = f_1(1) + f_2(1) + f_2(\sqrt{3})$.
3. $i=3, S = \{n_1, n_2, n_3, n_4, n_6\}, R = \{(n_1, n_2), (n_1, n_3), (n_2, n_4), (n_3, n_4), (n_4, n_6)\}$ and $P = 2f_1(1) + f_2(1) + f_2(\sqrt{3})$.

Due to the same reason in step 1, we should choose multi-path transmission route in this step. However, we choose the node 5 but not the node 7. From simple calculation, we can see that it is energy consuming if we reply the data packet from the node 7.

4. $i=4, S = \{n_1, n_2, n_3, n_4, n_5, n_6, n_8\}$
 $R = \{(n_1, n_2), (n_1, n_3), (n_2, n_4), (n_3, n_4), (n_4, n_6), (n_4, n_5), (n_5, n_8), (n_6, n_8)\}$
 $P = 2f_1(1) + 2f_2(1) + 2f_2(\sqrt{3})$.
5. $i=5, S = \{n_1, n_2, n_3, n_4, n_5, n_6, n_8, n_9\}$
 $R = \{(n_1, n_2), (n_1, n_3), (n_2, n_4), (n_3, n_4), (n_4, n_6), (n_4, n_5), (n_5, n_8), (n_6, n_8), (n_8, n_9)\}$
 $P = 3f_1(1) + 2f_2(1) + 2f_2(\sqrt{3})$

Finally, the algorithm stops here. The final result of the multi-path route topology is shown in Fig.4, where the solid lines indicate the bidirectional link with energy $f_1(d)$ and the dotted lines indicate the unidirectional link with energy $f_2(d)$.

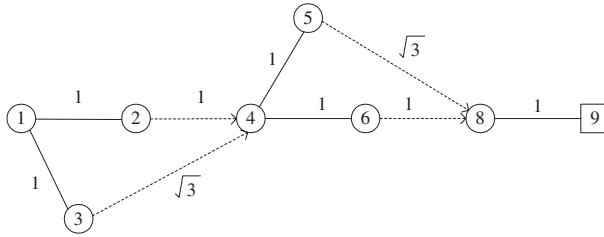


Fig. 4. Network topology of the example

4 Simulation and Analysis

We build up the MST by Kruskal’s algorithm of the wireless Ad Hoc networks and compare the energy consumption of the two cases below.

1. The route with only single transmission.
2. The multiple route with transmission diversity generated by our algorithm in Section 3.

In the simulation, we assume N nodes are randomly distributed in a $1000 \times 1000 \text{ m}^2$ simulation area. There is one monitor node fixed in the center of the simulation area. The transmission power function used are $f_1(d) = t_1 d^\alpha$ and $f_2(d) = t_2 d^\alpha$, where α is the constant between 2 and 4. In simplicity, assume the threshold is same for each node, therefore we can let $t_2 = 1$ and $t_1 = 6$. The network size N is defined as $10 \leq N \leq 1000$. We run 10 times with different seeds for each network size. The average results when $\alpha = 2$ for total energy consumption of each node are reported in Fig.5. Simulation is implemented by Glomosim [7].

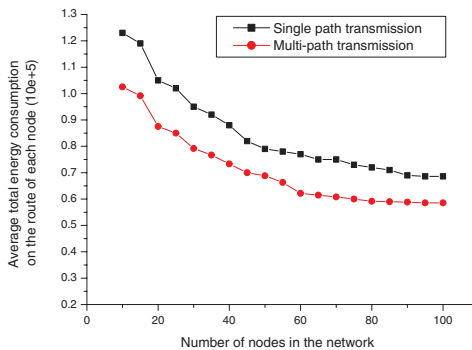


Fig. 5. Average total energy consumption on the route of each sensor

Fig.5 shows that the multi-path route with transmission diversity can save approximate 18% energy in same QoS (indicated by SER) compared with single path transmission. When the network becomes larger, the averaged total energy consumption on the route decreases slowly. It can be simply explained that there are two contradicting factors that contribute to energy consumption. They are network size N and edge length in the MST topology. For denser networks, the edge length is relatively shorter and the averaged energy on the route becomes smaller.

The simulation result reveals a fact that the multi-path route with transmission diversity can offer the same QoS as the single path transmission, but with low energy. This fact has another significant meaning. When the energy is going down, the nodes in the network can build up or partly build up multi-path route with transmission diversity to maintain stable QoS with less energy. That also answers how we make stable QoS in Wireless Multimedia Ad Hoc Network with Transmission Diversity.

5 Conclusion

In wireless mobile multimedia Ad Hoc networks, a new transmission scheme with physical transmission diversity can be employed to maintain stable QoS with lower transmission power of each node. Consequently, the routing topology will change from traditional single path to multi-path. In this paper, we construct an algorithm which can generate the corresponding multi-path route with transmission diversity in physical layer as well as network layer. The multi-path route has been proved of energy efficiency compared with traditional single path transmission. From the simulation results, we can see the average total energy of multi-path route is almost 18% less than single path transmission while the QoS is stably maintained in the same level (SER=1%).

References

1. G.N. Aggelou and R. Tafazolli, "QoS support in 4th generation mobile multimedia ad hoc networks", IEE Second International Conference on 3G Mobile Communication Technologies, Mar 2001.
2. D. GANESAN et al., "High-resilient, energy efficient multipath routing in wireless sensor networks", ACM SIGMOBILE Mob. Comput. Comm. Rev., Vol.5, No.4, 2001.
3. Xiaohua LI, "Energy Efficient Wireless Sensor Networks with Transmission Diversity", IEE Electronic Letters, Vol.39, No.24, Nov. 2003.
4. S.M. ALAMOUTI, "A Simple Transmitter diversity scheme for wireless communications", IEEE J. Sel. Areas Commun., Vol.16, No.8, pp.1451-1458, Oct. 1998.
5. T.S. Rappaport, "Wireless Communications: Principles and Practice", Prentice Hall, 1996.
6. Xiuzhen Cheng, B. Narahari, Rahul Simha, M. Cheng and D. Liu, "Strong Minimum Energy Topology in Wireless Sensor Networks: NP-Completeness and Heuristics", IEEE Trans. Mobile Computing, Vol.2, No.3, July 2003.
7. Glomosim page, <http://pcl.cs.ucla.edu/projects/glomosim>, 2003

Gaze from Motion: Towards Natural User Interfaces

Mun-Ho Jeong¹, Masamichi Ohsugi², Ryuji Funayama², and Hiroki Mori²

¹ Korea Institute of Science and Technology
mhjeong@kist.re.kr

² Toyota Motor Corporation
{ohsugi@nano,ryuji@funayamahiroki@mori}.tec.toyota.co.jp

Abstract. We propose a method of 3-D gaze estimation allowing the head motion under an uncalibrated monocular camera system. The paper describes the eyeball structure model with compact descriptions of the eyeball motion and its static 3-D structure. Assuming that the eyeball motion is independent of the head motion, we present a dynamic converging-connected model to make the gaze estimation allowing the head motion more systematic and simple. The gaze estimation is performed through the extended Kalman filter using the eyeball structure model and the dynamic converging-connected model. The preliminary test suggests satisfactory results.

Keywords: Gaze estimation, gaze from motion, eyeball structure, converging-connected model.

1 Introduction

Visual observation in human-machine interfaces is crucial to natural communication between men and machines. Gaze is not only a fundamental cue in the communication but also beneficial to other cues, face recognition, facial expression, gestures, etc.

Gaze estimation has been considered a difficult problem because gaze depends on not only the eyeball motion but also the head motion. In order to simplify the problem, therefore, the conventional methods adopted the assumption of stationary head and required the user to wear some equipment or to keep the head stationary [9][10]. It impaired the applicability and flexibility of gaze systems.

Recently, many non-intrusive approaches to gaze estimation have been studied. Baluja and Tan proposed a non-intrusive gaze tracking system using appearance-based methods [1][6], and showed results of high precision. There remained a problem of excessive training costs due to various head poses and different people.

The cornea-reflex model has been used widely in this area of non-intrusive methods [7][3]. The cornea-reflection-based systems give theoretical simplicity in estimating gaze since it does not have to consider the head motion, while

it requires the illumination of the environment to be carefully controlled to prevent undesired reflections in the iris. Instead of the cornea-reflex model Wang presented a method to estimate gaze by measuring changes in the iris contour [8]. As a rule, such big-eye approaches based on the precise measurements of the eye parts [4][5] could avoid the knotty problem to simultaneously consider both the eyeball and the head motions in gaze estimation, however, it takes expensive costs since they generally require a pan-tilt/zoom-in camera with sufficient resolution to accurately measure iris contour or pupil and additional stereo cameras for guiding the zoom-in camera to the eye position.

Matsumoto presented a real-time stereo vision system to estimate the gaze direction [13]. They could obtain the gaze direction by simple calculation, however they should be considerably careful to select two eye corners whose midpoint have to be on the line of the head direction starting at the eyeball center with known radius. Heinzmann determined the gaze vector from 3-D gaze direction relative to the head pose [15]. The key aspect of the method is that it uses a monocular camera to estimate the gaze under the head movements, but they still have to depend on the inner and outer eye corners.

In this paper we propose a novel method of non-intrusive gaze estimation allowing the head motion under an uncalibrated monocular camera system. The proposed method has similar aspects to SfM(structure from motion)[16] in a manner of estimating 3-D structure and motion. However, the proposed method is different from SfM in that it utilizes the dynamic 3-D structure including the eyeball motion in addition to the static 3-D structure of SfM and provides a specified dynamical model for the gaze estimation.

The dynamical model is based on independency between the head and the eyeball motions. The evaluation method of the estimated gaze is given, in which influences of two dominating factors on the gaze error is plotted. Then we are able to infer the accuracy of the estimated gaze using the plotted table and two factors from real experiments.

2 3-D Structure

Structure parameters of 3-D objects have been efficiently shown in Azarbayejani [16]. The pointwise structure only requires one parameter per point instead of three parameters. We adopt the pointwise representation to define the face structure and extend it to represent the eyeball structure.

2.1 Face Structure

The 3-D locations of the facial features with respect to the camera coordinates are represented as

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} u_i(1 + \beta\alpha_i) \\ v_i(1 + \beta\alpha_i) \\ \alpha_i \end{pmatrix}, \quad i = 1, \dots, N \quad (1)$$

where (u_i, v_i) is an image point of N facial points, β is the inverse of the camera focal length and α_i is the depth of each point. Thus, the structure of the face with N facial features is defined with the static unknown parameters $(\alpha_1, \alpha_2, \dots, \alpha_N)$ from the assumption that the face is a rigid object.

2.2 Eyeball Structure

The eyeball structure is expressed as the center points of both irises that undergo a change according to the eyeball motion. In order to model it, we introduce four coordinate systems that are the camera, the head, the eyeball and the gaze coordinate systems, where the Z-axes of the eyeball and the head coordinate systems are assumed to be parallel. We also assume that the head and the gaze coordinate systems coincide with the camera and the eyeball coordinate systems, respectively at the reference frame as shown in Fig. 1.

The rotation of the eyeball causes the position and orientation of the gaze coordinate system with respect to the head coordinate system to be specified by the chain product of the successive homogeneous transformations,

$$\mathbf{T}_h^g = \mathbf{T}_h^e \mathbf{T}_e^g = \begin{pmatrix} \mathbf{R}_{X_h, \pi} & \mathbf{t}_h \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{R}_e(\phi, \theta) & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}. \tag{2}$$

h : head coordinate system

e : eyeball coordinate system

g : gaze coordinate system

\mathbf{T}_m^n : homogeneous transformation from m to n

$\mathbf{R}_{X_h, \pi}$ is constant and denotes the rotation of π about the X_h -axis of the head coordinate system. \mathbf{t}_h is the position of the eyeball coordinate system with re-

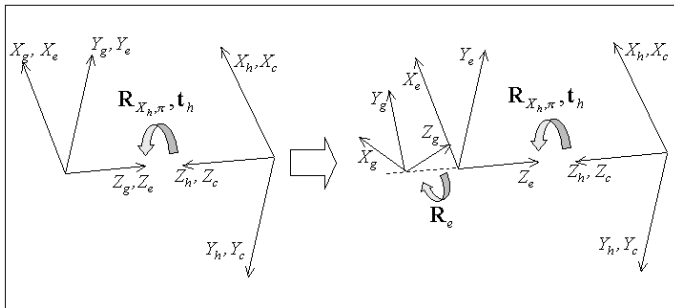


Fig. 1. Geometrical view on gaze. The left figure describes the coordinate systems at the reference frame while in the right one the gaze coordinate system moves with respect to the eyeball coordinate system by the eyeball rotation.

spect to the head coordinate system and is given as

$$\mathbf{t}_h = \begin{pmatrix} u_{ir}(1 + \beta\alpha_{ir}) \\ v_{ir}(1 + \beta\alpha_{ir}) \\ \alpha_{ir} + r_{eye} \end{pmatrix} \tag{3}$$

where r_{eye} is the radius of the eyeball, and (u_{ir}, v_{ir}) and α_{ir} are the image projection and depth of the iris center at the reference frame, respectively. $\mathbf{R}_e(\theta, \phi)$ has two degrees of freedom because of the eyeball’s rotationability and is shown as

$$\mathbf{R}_e(\phi, \theta) = \begin{pmatrix} \cos \theta & 0 & \sin \theta \\ -\sin \phi \sin \theta & \cos \phi & \sin \phi \cos \theta \\ -\cos \phi \sin \theta & -\sin \phi & \cos \phi \cos \theta \end{pmatrix} \tag{4}$$

to denotes the rotation of θ about Y_e -axis followed by the rotation of $-\phi$ about X_e -axis.

When the center of the iris moves from \mathbf{C}_0 to an arbitrary \mathbf{C} by rotating the eyeball as shown in Fig. 2, \mathbf{C} with respect to the camera coordinate system is obtained as

$$\begin{pmatrix} \mathbf{C} \\ 1 \end{pmatrix} = \begin{pmatrix} x_{ir} \\ y_{ir} \\ z_{ir} \\ 1 \end{pmatrix} = \mathbf{T}_c^h \mathbf{T}_h^g \begin{pmatrix} \mathbf{C}_0 \\ 1 \end{pmatrix} = \mathbf{T}_h^g \begin{pmatrix} 0 \\ 0 \\ r_{eye} \\ 1 \end{pmatrix} \tag{5}$$

where $\mathbf{T}_c^h = \mathbf{I}$ because the head coordinate system coincides with the camera one at the reference frame.

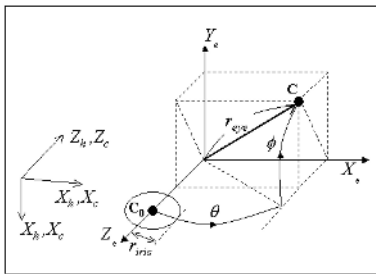


Fig. 2. Eyeball structure

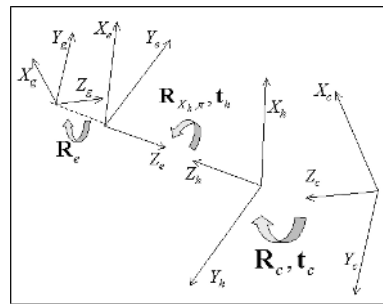


Fig. 3. Geometrical view on head motion

The projection of the iris radius at the reference frame is

$$s = \frac{r_{iris}}{1 + \beta\alpha_{ir}}, \tag{6}$$

and the radius of the eyeball is given as

$$r_{eye} = \frac{r_{eye}}{r_{iris}} \cdot \frac{r_{iris}}{s} \cdot s = \lambda s(1 + \beta\alpha_{ir}) \quad (7)$$

where λ is the ratio of the eyeball radius to the iris radius and generally considered constant.

Now we can define the eyeball structure by the position of the iris center with respect to the camera coordinate system. Substituting (2) and (7) into (5) gives

$$\mathbf{C} = \begin{pmatrix} x_{ir} \\ y_{ir} \\ z_{ir} \end{pmatrix} = \begin{pmatrix} (u_{ir} + \lambda s \sin \theta)(1 + \beta\alpha_{ir}) \\ (v_{ir} + \lambda s \sin \phi \cos \theta)(1 + \beta\alpha_{ir}) \\ \alpha_{ir} + \lambda s(1 - \cos \phi \cos \theta)(1 + \beta\alpha_{ir}) \end{pmatrix}. \quad (8)$$

(u_{ir}, v_{ir}) and s are measured at the reference frame and treated constant thereafter. The eyeball structure above has three independent terms except the inverse focal length β , which are the dynamic terms ϕ and θ related to the eyeball motion, and the static terms α_{ir} . It gives compact descriptions of the static 3-D structure and the eyeball motion.

3 Definition of Gaze

We define gaze as Z_g -axis' unit vector of the gaze coordinate system with respect to the camera coordinate system in Fig. 1. In order to obtain the gaze while the head is moving, we have to consider the eyeball motion and as well as the head motion. Under the fixed camera system, the head motion is described as the following rigid transformation as shown in Fig. 3,

$$\mathbf{T}_c^h = \begin{pmatrix} \mathbf{R}_c & \mathbf{t}_c \\ \mathbf{0} & 1 \end{pmatrix} \quad (9)$$

where \mathbf{R}_c and \mathbf{t}_c denote the rotation and the translation of the head coordinate system with respect to the camera coordinate system. Thus, the gaze with respect to the camera coordinate system is obtained as the third column vector of \mathcal{R} starting from \mathcal{P} in the following successive homogeneous transformation

$$\mathbf{T}_c^g = \mathbf{T}_c^h \mathbf{T}_h^g = \begin{pmatrix} \mathcal{R} & \mathcal{P} \\ \mathbf{0} & 1 \end{pmatrix}. \quad (10)$$

4 Dynamics Model

Feature points for gaze estimation are moved by not only the head motion but also the eyeball motion. This makes it difficult to track the feature points and thus to tackle the gaze estimation problem. For robust tracking of the features, it is necessary to consider a dynamical model which will be able to explain different patterns between the head and the eyeball motions. In this section we present an efficient dynamic model to track the head and the eyeball movements.

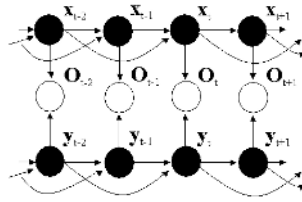


Fig. 4. Dynamic converging-connected model. The arrows denote probabilistic dependencies.

4.1 Dynamic Converging-Connected Model

On the statistical view the eyeball motion might be dependent on the head motion. But it is very difficult to find out the relationship since the dependency results from people’s identity as well as their emotional expressions. So it may be reasonable to take just a dynamical view on the motions.

Since the eyeball coordinate frame is fixed with respect to the head coordinate frame as in Fig. 3, we can assume that the eyeball motion is independent of the head motion. This idea and the fact that the eyeball has a different dynamic property from the head can be represented as the graphical model in Fig. 4.

The benefits of this model is that the number of dynamical parameters to be estimated by learning is reduced. This could lead to robust learning and thereby robust tracking. In the case that tracking is performed by the well-known particle filter [18] not by Kalman filtering approaches, the separation of states into two parts could reduce the required number of particles for robust tracking.

The state variable \mathbf{x} consists of the head motion and structure parameters like in [16]:

$$\mathbf{x} = (t_x, t_y, \beta t_z, w_{x_c}, w_{y_c}, w_{z_c}, \beta, \alpha_1, \dots, \alpha_{N+2})^T \tag{11}$$

where $(t_x, t_y, t_z)^T = \mathbf{t}_c$ and $(w_{x_c}, w_{y_c}, w_{z_c})$ are the location of the head coordinate frame and the interframe rotation with respect to the camera coordinate system, respectively, and $(\alpha_{N+1}, \alpha_{N+2})$ are the depth parameters of both iris centers. We assume that both eyeballs have the same motion with negligible error and compose the state variable \mathbf{y} using the dynamic terms in (8), that is,

$$\mathbf{y} = \begin{pmatrix} \phi \\ \theta \end{pmatrix}. \tag{12}$$

Under the assumption that the head motion is independent on the eyeball motion, the state variables, \mathbf{x} and \mathbf{y} , are probabilistically converging-connected given the observation vector.

The above graphical model can be specified as

$$\mathbf{x}_t = \mathbf{A}_1 \mathbf{x}_{t-1} + \mathbf{A}_2 \mathbf{x}_{t-2} + \mathbf{D}_x + \mathbf{u}_t, \quad \mathbf{u}_t \sim N(\mathbf{0}, \mathbf{Q}_x) \tag{13}$$

$$\mathbf{y}_t = \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \mathbf{D}_y + \mathbf{v}_t, \quad \mathbf{v}_t \sim N(\mathbf{0}, \mathbf{Q}_y) \tag{14}$$

$$\mathbf{O}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{y}_t) + \mathbf{w}_t, \quad \mathbf{w}_t \sim N(\mathbf{0}, \mathbf{R}), \quad t = 1, \dots, T \tag{15}$$

where \mathbf{A}_i and \mathbf{B}_i are the state transition matrices of each process, \mathbf{D}_x and \mathbf{D}_y are offset vectors, and $\mathbf{u}_t, \mathbf{v}_t$ and \mathbf{w}_t are the Gaussian random noises with the covariances, $\mathbf{Q}_x, \mathbf{Q}_y, \mathbf{R}$, respectively. The model parameters such as $\mathbf{A}_i, \mathbf{B}_i, \mathbf{D}_x, \mathbf{D}_y, \mathbf{Q}_x$ and \mathbf{Q}_y , are obtained by an EM learning [11][12]. The details are omitted.

4.2 Extended Kalman Filtering

The state space representation of the dynamic converging-connected model shown in (13-14) can be described by

$$X_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{pmatrix} = A_1 X_{t-1} + A_2 X_{t-2} + D + \mu_t \tag{16}$$

where

$$A_1 = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_1 \end{pmatrix}, A_2 = \begin{pmatrix} \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{pmatrix}, D = \begin{pmatrix} \mathbf{D}_x \\ \mathbf{D}_y \end{pmatrix}, v_t = \begin{pmatrix} \mathbf{u}_t \\ \mathbf{v}_t \end{pmatrix}.$$

And then it can be expressed more compactly by defining a state

$$\mathcal{X}_t = \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix}$$

and writing

$$\mathcal{X}_t = \mathcal{A}\mathcal{X}_{t-1} + \mathcal{D} + \mathcal{V}_t \tag{17}$$

where $\mathcal{A} = \begin{pmatrix} A_1 & A_2 \\ \mathbf{I} & \mathbf{0} \end{pmatrix}, \mathcal{D} = \begin{pmatrix} D \\ \mathbf{0} \end{pmatrix}, \mathcal{V}_t = \begin{pmatrix} v_t \\ \mathbf{0} \end{pmatrix}.$

The measurement process actually consists of the rigid transformation of 3-D model points defined in (1) and (8),

$$\begin{pmatrix} x_{c_i} \\ y_{c_i} \\ z_{c_i} \\ 1 \end{pmatrix} = \mathbf{T}_c^h \begin{pmatrix} x_i \\ y_i \\ z_i \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_c & \mathbf{t}_c \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ z_i \\ 1 \end{pmatrix} \tag{18}$$

and the projection of the transformed points onto the image plane,

$$\begin{aligned} \mathbf{o}_i &= \mathbf{f}_i(X_t) = \begin{pmatrix} \frac{x_{c_i}}{1+\beta z_{c_i}} \\ \frac{y_{c_i}}{1+\beta z_{c_i}} \end{pmatrix}, \quad i = 1, \dots, N + 2 \\ \mathbf{O}_t &= \mathbf{f}(\mathbf{H}\mathcal{X}_t) = \begin{pmatrix} \mathbf{f}_1(X_t) \\ \vdots \\ \mathbf{f}_{N+2}(X_t) \end{pmatrix}_t, \quad \mathbf{H} = (\mathbf{I} \ \mathbf{0}). \end{aligned} \tag{19}$$

The extended Kalman filtering is used to estimate the state because of the presence of nonlinearity in the measurement process. Since it is straightforward to perform the filtering given (17) and (19), we omit the explanation about it.



Fig. 5. Detection phase: from the left uppermost to the right lowermost.

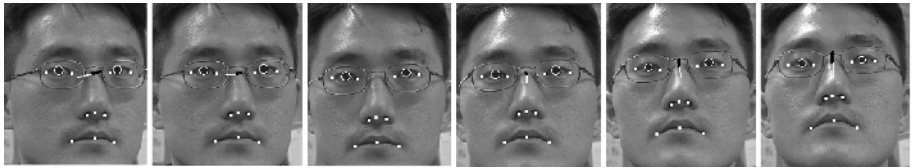


Fig. 6. Gaze tracking.

5 Implementation and Evaluation

The locations of the face and the facial features such as eyes, nose and mouth should be known before gaze tracking. We used the principal component analysis (PCA) to detect them [2]. In the PCA-based scheme, we constructed eigenfaces from the half-face images including both eyes and performed downsampling of input images. Then, the highly reduced computational costs enabled the detection phase to be implemented at the video-rate 30 fps. If the detected face keeps stationary for the predetermined frames, then the facial features are detected in the reduced searching area of the detected face as shown Fig. 5.

We set the frame when the facial features are detected initially to be the reference frame. And then the Hager tracker[14], which is an efficient region tracking method under affine transformation, begins to operate for tracking each feature. At the reference frame we selected two or three points from each feature region as the feature points, and constructed the 3-D structure models of the face and the eyeball using them as described in section 2.

A preliminary experiment with the 320x240 image sequences is shown in Fig. 6 where the black lines denote the head direction while the white ones show the gaze. The active contour method [17] was used to fit the projected iris contours. The current system of the gaze estimation works at 15 fps in a Pentium III PC of 933MHz.

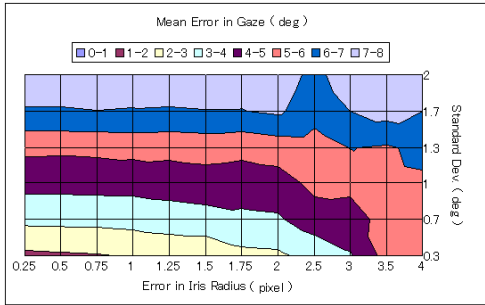


Fig. 7. Mean error.

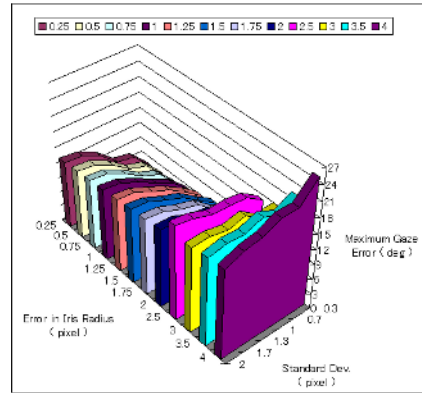


Fig. 8. Maximum error.

It is difficult to evaluate the accuracy of the gaze estimation system since we can seldom obtain the ground truth of the gaze. We used an indirect approach to evaluate it. There are two factors dominating the accuracy of the gaze estimation. One is the error in tracking the features. The other comes from the measurement error of the projected iris radius shown in (6).

We simulated their influences to the gaze error on the assumption that 320×240 image sequences are given. Fig. 7 shows the mean error of the gaze according to the measurement error of the iris radius and the standard deviation of the error in the feature positions. The maximum error in the gaze is described in Fig. 8.

As addressed in section 2.2, the iris radius measured at the first frame is used thereafter without correction, thus, the measurement error affect the gaze estimation adversely and continuously. However, from the Figures we found that the gaze error is nearly independent on the radius error if it is within 2 pixels. Actually, the radius error was measured below 1.3 pixels through a series of experiments.

In some experiments of the gaze estimation with 320×240 image sequences, the standard deviation of errors in the feature tracking was 1.5 pixels. This means that the error in the gaze estimation has 5-6 deg as its mean and the maximum 12 deg.

6 Conclusions

In the paper we presented a novel method of 3-D gaze estimation, in which we suggested an eyeball structure model with compact descriptions of the eyeball rotation and its static 3-D structure, and a dynamic converging-connected model to make the problem of gaze estimation allowing head motion more systematic and simple.

We performed some experiments and found that the proposed method is applicable to a real-time system. A simulation-based evaluation of the method showed that it has satisfactory accuracy in the gaze estimation.

References

1. S. Baluja & D. Pomerleau, Non-intrusive gaze tracking using artificial neural networks, TR CMU-CS-94-102, School of Computer Science, CMU, 1994.
2. M. Turk & A. Pentland, Eigenfaces for recognition, *J. of Cognitive Neuroscience*, Vol.3, No.1, pp.71-86, 1991.
3. ASL, Eye Tracking System Handbook, Applied Science Laboratories, Massachusetts, USA, 1996.
4. Kyung-Nam Kim, R.S. Ramakrishna, Vision-based Eye-Gaze Tracking for Human Computer Interface, *IEEE Int. Conf. on Systems, Man, and Cybernetics*, Vol.II, pp.324-329, 1999.
5. T. Ohno, N. Mukawa & S. Kawato, Just Blink Your Eyes: A Head-Free Gaze Tracking System, *Int. Conf. for Human-Computer Interaction*, Florida, USA, 2003
6. Kar-Han Tan, D.J. Kriegman & N. Ahuja, Appearance-based Eye Gaze Estimation, *IEEE Workshop on Applications of Computer Vision*, Orlando, USA, 2002.
7. K. Talmi & J. Liu, Eye and gaze tracking for visually controlled interactive stereoscopic displays, *Signal Processing: Image Communication* 14, pp.799-810, 1999.
8. J. Wang & E. Sung, Gaze determination via images of irises, *Image and Vision Computing*, Vol.19, No.12, pp.891-911, 2001.
9. K. Iwamoto & K. Tanie, Development of an eye movement tracking type head mounted display, *Proceedings of the 1997 IEEE International Conference on Robotics and Automation*, pp.2258-2263, 1997.
10. R. Jacob, The use of eye movements in human computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems*, 9(3):152-169, 1991.
11. Mun-Ho Jeong, Y. Kuno, N. Shimada & Y. Shirai, "Recognition of Two-Hand Gestures Using Coupled Switching Linear Model", *IEICE Trans. Inf. & Sys.*, Vol.E86D No.8, 2003.
12. Mun-Ho Jeong, Y. Kuno, N. Shimada, Y. Shirai, "Recognition of Shape-Changing Hand Gestures", *IEICE Trans. Inf. & Sys.*, Vol.E85-D, No.10, 2002.
13. Y. Matsumoto & A. Zelinsky, An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement, *Proceedings of IEEE fourth Int. conf. on Face and Gesture Recognition*, pp.499-505, 2000.
14. G.D. Hager & P.N. Belhumeur, Efficient Region tracking with parametric models of geometry and illumination, *IEEE Trans. PAMI*, Vol.20, No.10, Oct. 1998.
15. J. Heinzmann & A. Zelinsky, 3-D facial pose and gaze point estimation using a robust real-time tracking paradigm, *IEEE Int. Workshop on Automatic Face and Gesture Recognition*, pp.142-147, 1998.
16. A. Azarbayejani & A. Pentland, Recursive estimation of motion, structure, and focal length, *Perceptual Computing TR-243*, MIT Media Lab., 1994.
17. A. Blake & M. Isard, *Active Contours*, Springer-Verlag, 1998.
18. Isard, M. and Blake, A., "Condensation-Conditional Density Propagation for Visual Tracking", *Int. J. Computer Vision*, Vol.29(1), pp.5-28, 1998.

The Development of MPEG-4 Based RTSP System for Mobile Multimedia Streaming Services

Sangeun Lee, Hyunwoo Park, and Taesoo Yun

Dept. of Digital Virtual Reality, Division of Digital Contents, Dongseo University, Sasang-Ku,
Busan, South Korea.
sangeun22000@hotmail.com, {phw1010, tsyun}@dongseo.ac.kr

Abstract. This paper describes the implementation of IOCP based a MPEG-4 RTSP(Real Time Streaming Protocol) system for mobile environment. To prevent the status of cut-off and delay due to a number of contemporary users, we designed and implemented IOCP(I/O Completion Port) based MMS system which supports user access to the internal structure of RTSP through fast and small thread to enhance the performance of RTSP socket connection. This paper provides portability of MPEG-4 FGS coding scheme into mobile environment such CDMA based internet.

1 Introduction

Recently, the transmission of multimedia content such as VOD(Video On Demand) service over the Internet(WWW) has been growing steadily over the past few years[1]. For the purpose of better multimedia service over internet, we should consider the limit of network bandwidth and the characteristics of service according to interaction among the users. Therefore, MPEG-4 based streaming service development demanding low bandwidth and high quality of video become active.

Especially, according to the speed-up of wireless communication and acceleration of miniaturization, lightness, and low electric power of terminal, the applications for the traditional communication by wire are moved into the wireless applications such as mobile VOD or mobile MMS services. However, since the mobile environment was designed for low data communication, satisfying the necessary requirements for the effective delivery of multimedia streams poses significant challenges. For examples, the mobile environment is characterized by large bandwidth variations due to heterogeneous access-technologies of the receivers, e.g., analog modem, cable modem, CDMA, etc.) or due to dynamic changes in network conditions, e.g., congestion events[2]. In previous works, to solve these problem, e.g., the variation of bandwidth, packet loss during sending, etc., MPEG-4 FGS or FGST encoding schemes are adopted.

In terms of processor, if the contemporary users are large, the problem deepen the status of cut off and delaying. To cope with these problems, in this paper, we designed and implemented IOCP(I/O Completion Port) based MMS system which support user access to the internal structure of RTSP through fast and small thread to solve this

problem. Therefore, our system enhanced the performance of RTSP socket connection. The RTSP module is designed to process a large number of socket connection using work thread of 64 and one Scavenger thread.

Especially, mobile streaming technologies implemented in this research such as FGS and IOCP based server structure are likely to be used as core technology in VOD(Video On Demand) or MMS(Multimedia Messaging Service).

The paper is organized as follows, In Section 2, the Overall system of RTSP streaming system implemented in this paper is described. Subsequently, in section 3, experimental results and evaluation of this system. Lastly, section 4 describes the conclusions and future works.

2 The Overall System

The overall structure of our RTSP(Real-time Streaming Protocol) System is displayed in Fig. 1. The system is consists of three parts: MPEG-4 FGS(Fine Granular Scalability) encoder, RTSP Server and RTSP Client. The MPEG-4 FGS encoder compresses the input image obtained by CCD or other typed Camera. The RTSP server send the video compressed by FGS encoder over the network in stream, and the RTSP client is the part that play the video combining and reconstructing the received multimedia stream slice called VOP(Video Object Plane).

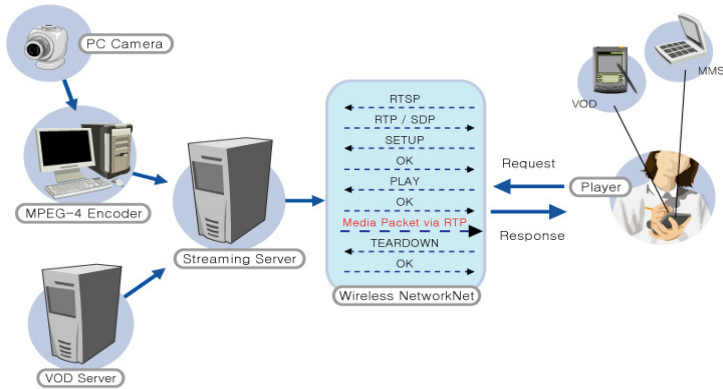


Fig. 1. The overall structure of RTSP system

2.1 The Real-Time Codec

We use MPEG-4 FGS encoding scheme adopted in MPEG-4 standard to compress the image captured by the camera. The MPEG-4 FGS scheme encode a series of image to basic layer and enhancement layer[3,4]. The FGS scheme consists of two method, i.e. DCT bit-plane encoding scheme and wavelet encoding scheme. In this paper, we

use bit-plane encoding scheme based on DCT transformation. As shown in Fig. 2. FGS scheme consists of two layer. The basic layer encoded appropriate to the current network status demanding the lowest bound of bit rate. The enhancement layer which provides the very effective and simple process is the one uses fine-granular encoding scheme at the maximum bit rate.

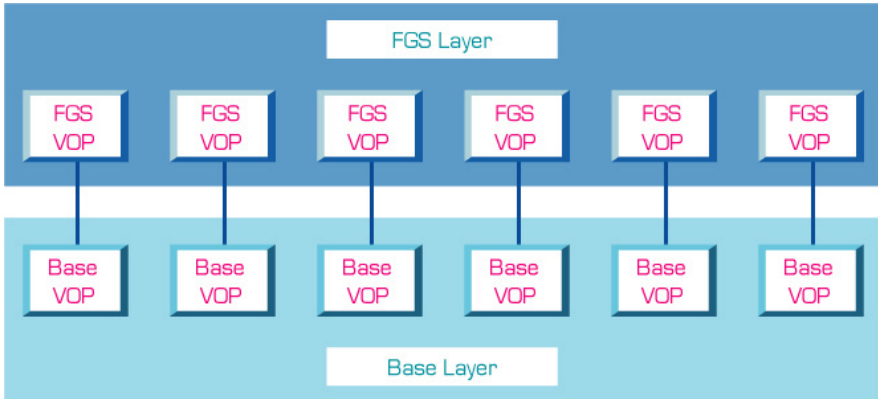


Fig. 2. The hierarchical structure of MPEG-4 FGS.

The base layer uses non-scalable coding to reach the lower bound of the bit-rate range and is similar to the traditional decoder. The enhancement layer is to code the difference between the original picture and the reconstructed picture using bit-plane coding of the DCT coefficients[4].

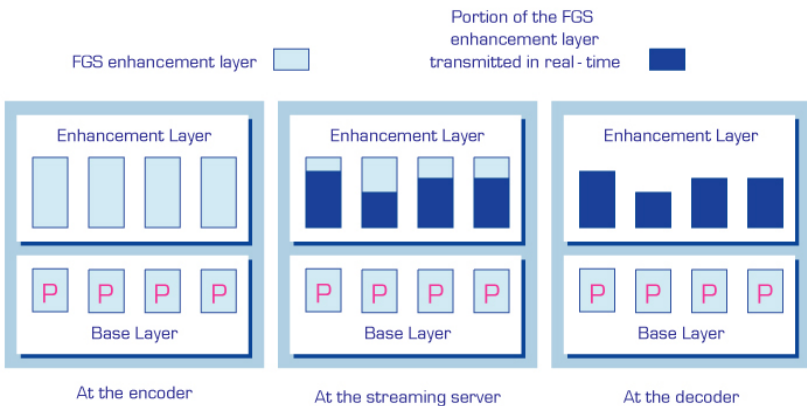


Fig. 3. The Internal structure of FGS System.

Fig. 3. display the internal structure which MPEG-4 FGS encoder encode the image obtained by CCD Camera in real-time and send via the server. The size of encoded image is QCIF(176 x 144) which is appropriate mobile environment.

2.2 FGS(Fine Granularity Scalability) Structure

To send video sequence over the internet which is various bandwidth, high rate of packet loss, the system with high error tolerance and adaptive to the various bandwidth is requisite. For the purpose of that, MPEG-4 of ISO/IEC select FGS(Fine Granular Scalability) to the standard.

The basic idea of FGS is to code a video sequence into a base layer and an enhancement layer. The base layer uses non-scalable coding to reach the lower bound of the bit-rate range and is similar to the traditional decoder. The enhancement layer is to code the difference between the original picture and the reconstructed picture using bit-plane coding of the DCT coefficients[4].

Fig. 4 and 5 show the FGS encoder and the decoder structures, respectively. The bitstream of the FGS enhancement layer may be divided into any number of bits appropriate to each picture after encoding is completed. As shown in Fig, FGS can provide various quality of picture for each VOP(Video Object Plane). The enhancement layer constitute VOPs enhance the VOPs of base layer[4].

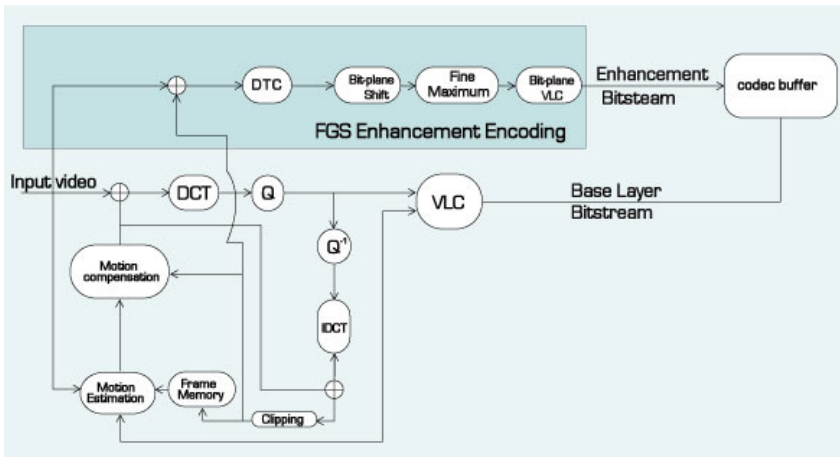


Fig. 4. The Encoder Structure

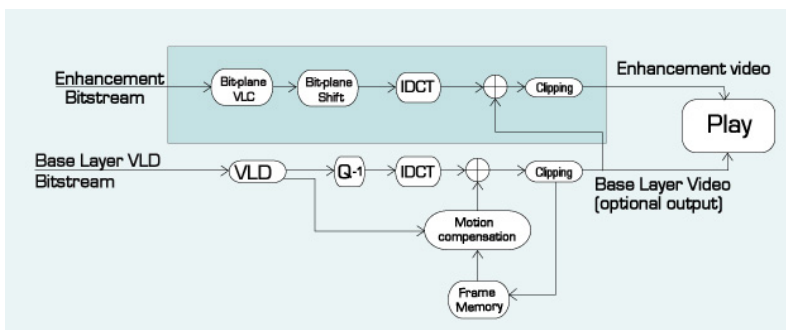


Fig. 5. The Decoder Structure

The decoder should be able to reconstruct an enhancement video from the base layer and the truncated enhancement layer bitstreams. The enhancement-layer video quality is proportional to the number of bits decoded by the decoder for each picture.

2.3 RTSP(Real-Time Streaming Protocol) Server / Client

The RTSP Server has the role of transferring the multimedia data encoded in MPEG-4 encoder to the terminal(clients). Both TCP/IP and UDP protocol which is used in IP based packet communication environment is not adequate if we consider the characteristics of streaming services. In this research, we used RTSP(Real-time Streaming Protocol) proposed by IETF to transfer multimedia with real-time variability[5,6]. The RTSP generates and controls the streams synchronized temporally such as audio and video. But it generally don't send successive-media itself, remote control the network for multimedia server[6].

The difference between protocol for general multimedia application and RTSP is that RTSP controls all of the function required server and terminal such as the setting of streaming session, start and end. The RTSP Server performs the streaming service that send multimedia data to clients after analyzing and process the request of clients. The RTP based on UDP has the several information, i.e. the number to represents the sequences of packets, the time stamp, the types of media.

RTSP Client implemented in this paper minimizes the initial delay of streams transferred from server and execute the function of playing streams. At this time, it should analyze QoS(Quality of Service) information using header information of RTP packet and feed back several to the server.

3 The IOCP (I/O Completion Port) Based Server

IO Completion ports is a method Windows uses to notify and awaken a thread when asynchronous IO is complete. This is more efficient than creating one thread per IO operation and having that thread wait for the operation to complete. Instead, a pool of threads is created (in this case two threads per processor). These pool of threads block on the IO completion port and when an IO operation is complete one thread wakes up and performs another operation. This will benefit servers with hundreds to thousands of players. It does not benefit games with only a few players and in fact may be slower due to Windows overhead.

Therefore, multiple instances of servers and clients on one computer will all share the worker threads. This improves efficiency in cases such as running multiple servers on one machine, or a client and a server on the same machine.

In this paper, we modify the traditional server architecture with new server algorithm, i.e., IOCP(I/O Completion Port) and therefore, we can enhance the performance of socket connection. Fig.6. displays the IOCP architecture used in our RTSP streaming system.

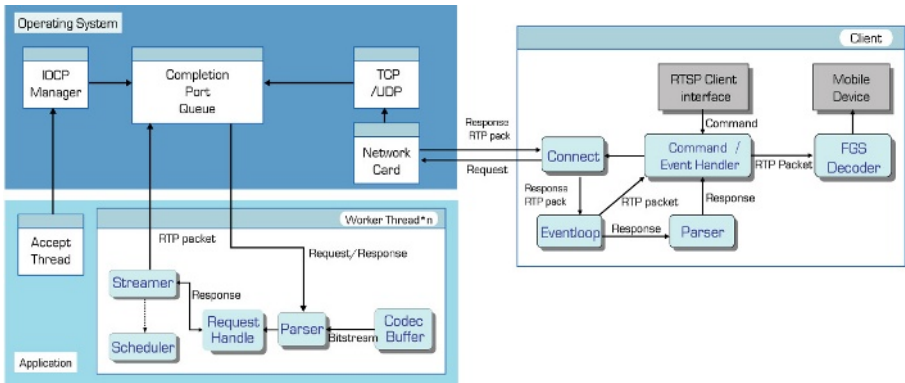


Fig. 6. The Architecture of RTSP Server and Client base on IOCP

4 The Experimental Results

To verify the performance of our MPEG-4 based RTSP system, we implemented the MMS Server/Clients using FGS encoding scheme and tested wireless PDA streaming service. The experimental environment is composed of Pentium IV with 1.0Ghz, 512M Ram, and wireless PDA with COMPAQ SA1110 200Mhz, 64MB Ram, and smile cam. We use Microsoft Visual C++ 6.0 and Embedded Visual C++ 3.0 to develop the server and client, respectively. We also developed the PDA emulator customized to our Desktop environment to display streaming status. The PDA emulator and desktop GUI is displayed in Fig. 7.



Fig. 7. PDA Emulator and GUI for Server

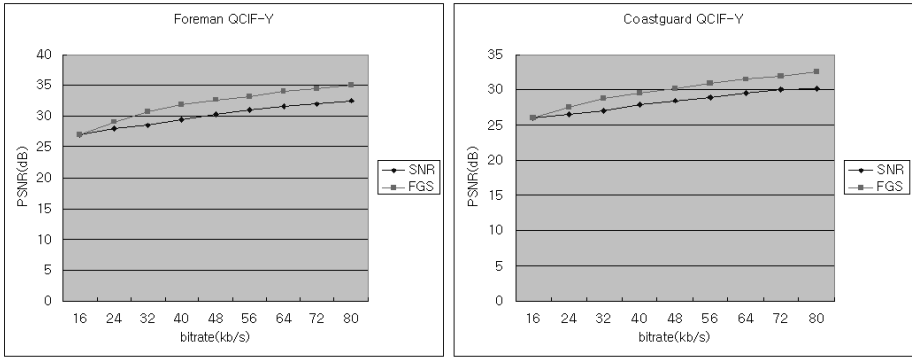


Fig. 8. Performance of FGS coding and traditional MPEG-4 SNR coding .



Fig. 9. Mobile Streaming and VOD Service

We also compared the effectiveness of MPEG-4 FGS between general SNR scalability video coding to verify our system. The test image used in the experiment was Foreman QCIF(176 x 144). The PSNR was calculated in the bit rate of 16, 24, 32, 40, 48, 56, 64, 72, 80kbps and the result was displayed in Fig.8. As shown in Fig.8, the FGS encoding scheme is superior to the SNR video coding from 1 to 2db in average.

We compress the images through the CCD Camera into the size of 176 x 144, QCIF image, and service to Compaq PDA(Personal Device Assistant) with the performance of 7frame per second. But the streaming speed is very low due to the performance of CPU. We also understood the fact that streaming speed is up to 10frame / second when we use subQCIF(128x96) appropriated CDMA network.

5 Conclusions and Future Works

In this paper, we implemented the MPEG-4 based RTSP system for the mobile multimedia streaming service. We implemented streaming system using FGS encoding scheme to sustain the robustness to the variation of network bandwidth.

In terms of processor, if the contemporary users are very many, The problem that is deepen the status of cut off and delaying. In this paper, we designed and implemented IOCP(I/O Completion Port) based MMS system which support user access to the internal structure of RTSP through fast and small thread to solve this problem. Therefore, our system enhanced the performance of RTSP socket connection.

Especially, mobile streaming technologies implemented in this research such as FGS and IOCP based server structure are likely to be used as core technology in VOD(Video ON Demand) or MMS(Multimedia Messaging Service).

References

- [1] H.M. Radha, M.V.D. Schaar and Y. Chen, "The MPEG-4 Fine-Grained Scalable Video Coding Method for Multimeia Streaming Over IP," IEEE Trans. On Multimedia, Vol. 3, No. 1, March, 2001, pp, 53 ~68.
- [2] M.V.D. Schaar and H.Radha, "A Hybrid Temporal-SNR Fine-Granular Scalability for Internet Video," IEEE Trans. On Circuits and Systems for Video Technology, Vol. 11, No. 3, Mar. 2001, pp 318-331.
- [3] Coding of Audio-Visual Objects, Part-2 Visual, Amendment 4: Streaming Video Profile, ISO/IEC 14496-2:1999/FDAM, October, 2000.
- [4] Weiping Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard" Circuits and Systems for Video Technology,IEEE Transactions on,Volume: 11Issue: 3, March 2001.
- [5] H.Schulzrinne, A. Rao, R.Lanphier, "Real Time Streaming Protocol(RTSP)", RFC-2326, 1998
- [6] Jung-Gu Kang, Tea-Uk Choi, Young-Ju Kim, Ki-Dong Chung, "Implementation of RTSP Server and Client for Multimedia Streaming int the Internet", The Proc. Of KIS, Vol, 06 NO. 06, pp. 1~4, 1999.

Seamless Mobile Service for Pervasive Multimedia

Enyi Chen, Degan Zhang, Yuanchun Shi, and Guangyou Xu

Key Lab of Pervasive Computing
Dept. of Computer Science and Technology
Tsinghua University, Beijing, 100084, P.R.China
{chenenyi00@mails, gandegande@mail}.tsinghua.edu.cn
<http://media.cs.tsinghua.edu.cn/~pervasive>

Abstract. In this paper, we propose a manager of seamless mobile service for pervasive/ubiquitous multimedia, which can dynamically follow the user from place to place without user awareness or intervention by layering architecture of component platform and agent-based migrating mechanism under consideration of robustness, scalability and load balancing. The manager can custom multimedia task and encode service descriptions by the XML or SMIL technology to provide flexibility, it can also discover the services, filter, synthesize or handoff them according to the context information and match between the task and service, especially, it can manage seamless mobility of pervasive multimedia. The validity evaluation of the manager has been done by experimental demo.

1 Introduction

How does a mobile device utilize available resource in the surrounding for service advertisement, discovery, filtration, synthesis and migration? With the shift of the history and context of pervasive multimedia during the mobility of user or task of user, how does the computing device and software resource around it make adaptable change for seamless mobility [1] [2]. How to deal with the problem of seamless mobility and realize the transparent transferring of task is one of key technologies in pervasive computing. Currently, the useable technology is like mobile IP used as network-level protocol, fixed or mobile agent used in application-level cases. Migration based on mobile agent is one kind of important method in pervasive computing paradigm. The focus is how to automatically manage and coordinate useable distributed computing environment, record its history, and restore its context seamlessly so as to continue the task and meet the requirement of mobile application. The chief function requirement of seamless mobility is focused on the continuity and adaptability of pervasive multimedia [3]. The continuity is that the application can pause and continue to work later without the loss of the current state and the running history. The adaptability is that the application is not restricted by computing device and context of service but adaptable to its environment. In our opinion, this is a kind of mobile working paradigm [4]. But when seamless migration for computing task of pervasive

multimedia is realized on PC, laptop, or PDA, there are many difficult problems to be solved [5] [6]. The rest of this paper will be organized as follows. Firstly, we give the Scenarios and architecture of manager for seamless mobile service, then we describe mapping between task and useable service, design a kind of strategy of seamless mobile service, including solving the following several sub-problems: avoiding migration failure and remainder dependency. Finally, we evaluate the validity of our manager for seamless mobile service and draw a conclusion.

2 Scenarios

Scenario I: When you fly to Beijing Capital International Airport from Kyoto Airport, you may transfer in Tokyo International Airport. It is very boring since you have to wait for another two hours before your next flight in the lounge. So you use your wireless enabled PDA to search for “games”. You find another traveller who launches a “Chinese Chess” game, which is just your favorite. So you connect to his device and play with him.

Scenario II: When you work in front of your personal computer (PC), your mobile phone rings. After you pick it up, what make you surprised is that the caller’s live video is displayed on your computer screen though the user’s mobile phone can display and capture the live video. Vice versa, your live video is also displayed on the caller’s computer screen.

3 Architecture of Manager for Seamless Mobile Service

The architecture of our seamless migration manager is also Multi Agents System (MAS) including fixed or mobile agents. It can run on networks connected PC, laptop, or PDA. Now we introduce the manager. The runtime environment is composed of four kinds of components: Human-computer interaction interface, Task manager, Continuity manager and Service manager (as the figure 1). The introduction of them is as followings:

1) *Human-computer interaction interface* including agent interface and relative controlling. The agent interface is used as defining the attributes of agent, such as ID of agent, Name of agent, Type of agent (such as TA, SA, UA, VA, EA, DA, and so on), Current status of agent (one of five status are “Ready”,

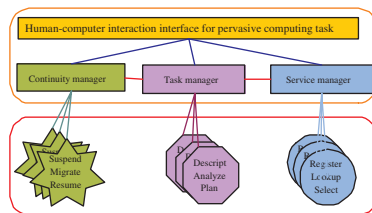


Fig. 1. Components of our seamless migration manager

“Waiting”, “Transferring”, “executing”, “Dead” or “Destroyed”), Association relationship of agent (including relationship between agent and task of learning, relationship between agent and agent).

2) *Task Manager* is for application service, which manages the application/task array, including task description of learning, task analyzing, mapping or binding between task and service, loading, executing, planning schedule of task of learning, etc.

3) *Continuity Manager* is for preparing “Migrating travel plan/schedule” of task of learning, sensing context, suspending of task, historical status (including log, configuration, etc.) recording, agent management, addressing of target node, determining of transferring granularity which is for avoiding the transferring failure, reducing the remainder dependency & contracting the transferring delay, resume of task of learning, and so on.

4) *Service Manager*. It manages the registration of service, service discovery [8], lookup of service, service selection/association, and mapping or binding between task and service.

These components can communicate each other, and may be controlled by human-computer application interaction interface including agents and relative control, which is individual interface for PC, laptop, PDA.

4 Adapted Mapping Between Task and Useable Service

Figure 2 shows adapted mapping between task described with XML or SMIL and useable service, where the transaction T wanted to be completed by user is named Task, which constitutes of subtask or sub-transaction T_i , each T_i is independent unit of function. Because of the diversity of task, its subtask is different each other, in order to keep their compatibility, the description of subtask should be abstract, mainly, the key and necessary parameters. The execution of task is finished by the service supplied by relative resources R, but different subtask, the done method is different too, we can be classified it into three kinds: Event Type, Stream Type, Bulk Type.

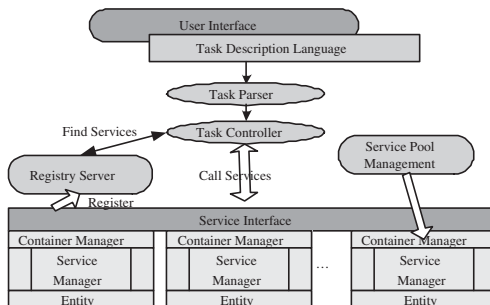


Fig. 2. Adapted mapping between task and useable service

Tasks can be described normalizedly by XML or SIML, the format by XML is as follows:

```
<?xml version="1.0"?> <Task_Descriptor> <head><name>TaskX
</name> <comment></comment></head><body><Task_Owner></Task_Owner>
<Service_Attribute></Service_Attribute> </body></Task_Descriptor>
```

5 Strategy of Seamless Mobility Based on Agent

As a kind of special computing resource, agent supports deployment of computing resource and mobility freely, which makes the system manage and adjust easily, so it is suitable for application of seamless mobility.

If the transferred amount of data is partial, and this part of information must be transferred firstly so that the task can restore the runtime environment and run continuously on the target node, this part of information is regarded as “Key Set”, so we can divide the migrated information into several chunks, such as executing code chunk, running status chunk, and so on. For resuming, the “Key Set” chunks must be integrated, otherwise it is impossible to go on running continuously.

According to classification of agent in our test bed, we make the following rules:

Navigation Agent (VA) need NOT do direct relative works with the task, which is familiar with the topological structure of Internet/subnet of target node and addressing in the network. The Data Structure of VA may be divided into two parts: one is itself “function body”, another is MessageBox (MB, mark as) used as loading moved object and transferring in the Internet/subnet node. Task Agent (WA) does detail jobs, which includes executing the code, managing the data and environmental status, and so on. It can transfer with the Navigation Agent (VA) in the network and need NOT know the structure of Internet/subnet. When migrated, WA seeks relative VA and joins in its MB firstly, then sent to target node by VA.

The designed and adopted algorithm of seamless migration by us is as follows:

1) According to the prepared TRAVEL SCHEDULE/PLAN for migrating / transferring, logic node PA2 lets VA begin addressing in the network according to the address supplied by logic node PA3, when the connection is successful, VA sends instruction “TransferNode” to Logic node PA3 as target node, VA+WA transfer to PA3 after packing. Logic node PA3 sends instruction “UpdateLinking” to all logic nodes connected to PA2, such as logic node PA1.

2) When PA1 has received the instruction “UpdateLinking”, it creates the association to new link, and sends instruction “LinksUpdated” to logic node PA2. When PA2 has received all expected message “LinkUpdated”, and then sends instruction “ActiveNode” to logic node PA3. According to the topological relationship, under the rule of FIFO, PA3 activates the Messenger, up to now, the transferring work is finished. When all Messengers are activated, each WA will restore running environment and do instruction “ExecuteTask”. During the

executing of each WA, on the one hand, the historical snapshots will be recorded and saved, on the other hand, VA do the instruction “ListenTask” continuously and get the next transferring instruction “TransferSignal”. During the executing of WA, VA checks the prepared TRAVEL SCHEDULE/PLAN, if another new migrating plan is checked, the “TransferSignal” instruction will be sent to WA. If no instruction “TransferSignal” is received by WA, it will execute its task continuously until the task is completed, otherwise, it will stop executing, and Goto 1 for the new migrating/ transferring.

6 Focus of Seamless Mobile Service of Multimedia

6.1 The Migration Failure Problem

In the pervasive computing environment, because the position of agent is often variable, the cases may be occur that when the agent1 is being transferred to agent2 and embedded in it to do the task together from node C, but during the transferring, the agent2 has moved from node C to node D, when agent1 arrives at node C, it can NOT find the agent2. This case is called the migration failure problem. This kind of problem can NOT keep the continuity of transferring of task.

In order to solve this problem, we think there are three factors should be considered: 1) When the position of agent has moved, how to know this change by other relative agents. 2) When the transferring of agent, how to deal with the message sent to it. 3) During the transferring of agent, whether the receiving agent can be transferred freely or not.

Our solution is as follows: 1) When the agent moves to new node, it should send “Notation” Message to all other relative agents 2) Before the agent prepares to be transferred, it should query the current position of receiving agent, at the same time, the transferring relationship and event should be sent to it by message 3) Determine the transferring topological relationship of agent, based on the rule “FIFO”, when each agent begins to be transferred, the receiving agent may be transferred but a little later should be locked, after the transferring process is over, and the receiving agent should be unlocked. For it, a signal semaphore may be set.

In our test bed, the “Notation” message may adopt three kinds: Unicast, Multicast, Broadcast. The synchronic mechanism “addressing first, then locking and transmitting” designed based on the “time-topological” relationship by us can realize the synchronization between transferring agent and receiving agent, the transferring failure problem can be solved radically and adapted for all kinds of application pattern, besides avoiding the failure, the restriction to receiving agent may be reduced, so it is general. Where, “time-topological” relationship may be considered in the “Travel Plan/Schedule for transferring”. The schedule may consist of certain travel sequence, the data structure of each travel sequence is like Table 1.

In Table 1, TP_MC, TP_SP, TP_RE are important for basic migrating / transferring operation, that is to say, Only the TP_MC is OK, “Travel Plan for

Table 1. Data structure of each sequence

No	ID	Description	No	ID	Description
1	TP_ID	ID of plan	7	TO_IPC	Source IP of migrated object
2	TP_NP	Name of plan	8	TO_IPD	Target IP for migrated object
3	TP_MD	Made date of plan	9	TP_MC	Migration condition
4	TP_ON	Order number of plan	10	TP_SP	Snapshot point of suspension
5	TP_NO	Name of migrated object	11	TP_RE	Resume point of task
6	TP_MG	Migrated granularity	12	TP_RI	Running ID after resuming

transferring” may be run, meanwhile, record and save “TP_SP” and “TP_RE”, both for restoring the running environment.

Whether TP_MC is OK or not, the following aspects should be checked: Whether the current status (may be one of five kinds: Ready 1, Waiting 2, Transferring 3, Executing 4, Destroyed 5) of agent is Waiting 2, whether the target address TO_IPD may be reached or not, whether the threshold of transferring delay is OK or not, whether the reminder dependency cases may exist or not.

6.2 The Remainder Dependency Problem During Migrating

In fact, there are three valid mapping forms based on the three factors above:

1) Strong Transfer, suspends after transferring, resume/restore and run after finishing the whole information

2) Strong Transfer, suspends after transferring, resume / restore and run timely after finishing the Key Set information (at the same time, the whole remainder information will continue its transmitting)

3) Weak Transfer, suspends after transferring, resume/restore and run after finishing the Key Set information (at the same time, the selected partially information will continue transmitting to the target node)

In the first mode, the transferring delay of agent includes that packing the whole information and transmitting, restoring the agent and the whole information on the target node; In the second mode, includes that packing the Key Set information and transmitting, restoring the agent and the Key Set; In the third mode, includes that packing the Key Set information and transmitting, restoring the agent and the Key Set. In the same condition, the delay of the first one is the longest, the third one in the shortest. In the second and third migrating/transferring mode, because of selecting a part information as the “Key Set” and being transferred firstly, but for different application, it is NOT known that which part of information is necessary, a certain “Key Set” may NOT migrate to the target node timely, the running agent must wait for it, that is to say, running agent is still dependent on part of information on the source node, this case is called “remainder dependency”. This problem may lengthen the transferring delay of task, when it is serious, it will influence the seamless mobility, so this

case must be avoided. In our opinion, the “remainder dependency” problem will be solved from two aspects:

1) Tuning reasonably the transferring granularity of task. We divide it into several parts, and they deal with by their relative agent, such as Execution-code Agent (EA), Data Agent (DA) and other Agents (such as Environment-state Agent). If too larger, it is restricted by the bandwidth; If too smaller, transferring time is much more. Both may lengthen the delay. Based on analyzing on theory and application tests, we adopt a kind of partition method called “subsection” or “pagination”, the size or number of “section” or “page” is determined adaptively by bandwidth, cache buffer, volume of MessageBox (MB). When this case is occurred, the necessary information may be transmitted through “section interruption” or “page interruption”, but the frequency should be adjusted adaptively according to historical record information.

2) Optimizing the Key Set. The Key Set will be determined adaptively according to nearest principle and used frequently principle and cut off the redundancy information. The relative adapted strategy may be referred the bibliography [9].

7 Evaluation of the Manager

Currently, we supply the function that the task dynamically follows the user from place to place and machine to machine. For example, the video-playing task may follow me from my house to other places, such as my office, stadium, coffee house, park, airport, etc., and vice versa. The task is described partially by SMIL in the following. Because the migrating/transferring mode based on agent is distributed or peer-to-peer / end-to-end, we have designed several relative communication primaries for migrating, for example,

1) BeginToListen Primary.

```
BeginToListen(UINT nPort, ACCEPT_CALLBACK callback);
Void(CAgent::*ACCEPT_CALLBACK)(UINT &Connection_ID);
```

2) BeginToRequest Primary.

```
BeginToRequest(UINT &nConnection_ID, CString IP,UINT nPort);
```

3) Transfer Primary.

```
Transfer(UINT nConnection_ID, CString strMsg, CDate time_stamp);
```

Based on the above test bed, one snapshot comparison result of experiments for seamless migration from PC to PC, laptop and PDA is shown. From the results, the delay time from PC to PC is the shortest under the same evaluation framework, but the delay time from PC to PDA is the longest.

Of course, Security is a big problem for test bed. We will provide security by applying existing forms of public-key cryptography and Intrusion Detection Agent (IDA). During seamless migration, the intrusion cases include illegal access, hostile attack through all kinds of invalid channel or approach, etc, so the detection mechanism to be proposed by us includes monitoring, auditing, analyzing, warning, response, and so on [6].

8 Conclusions

Based on the application requirements of pervasive computing, we have proposed a seamless migration manager for pervasive multimedia, which supplies the function that the task dynamically follows the user from place to place. Our key insight is that this capability can be achieved by layering architecture of component platform and agent-based migrating mechanism. In this paper, we have given the architecture of manager for seamless mobile service, described mapping between task and useable service, designed a kind of mechanism of seamless migration, including solving these problems, such as method of seamless migrating, avoiding migration failure and remainder dependency. The validity of the manager has been evaluated by the demo.

References

1. M. Satyanarayanan, Pervasive Computing: Vision and Challenges, IEEE Personal Communications, vol.8, no.4, pp. 10-17, August, 2001
2. M. Kozuch and M. Satyanarayanan, Internet Suspend/Resume, In Proceedings of Fourth IEEE Workshop on Mobile Computing Systems and Applications, Calicoon, N.Y., Jun. 1, 2002
3. K. Takasugi, Seamless Service Platform for Following a User's Movement in a Dynamic Network Environment, In Proceedings of the First IEEE International Conference on Pervasive Computing and Communications (PerCom'03), Fort Worth, Texas, USA, March 23-26, 2003
4. Y.C. Shi, W.K. Xie, G.Y. Xu, R.T. Shi, E.Y. Chen, Y.H. Mao and F. Liu, The smart classroom: merging technologies for seamless tele-education, IEEE Pervasive Computing, vol.2, no.2, pp. 47-55, April-June, 2003
5. E.Y. Chen, Y.C. Shi, D.G. Zhang and G.Y. Xu, A Programming Framework for Service Association in Ubiquitous Computing Environments, In Proceedings of 4th IEEE Pacific-Rim Conference on Multimedia(PCM2003), vol.1, IEEE Press, pp. 202-207, Singapore, December 15-18, 2003
6. D.G. Zhang, G.Y. Xu, Y.C. Shi and E.Y. Chen. Mobile agents with intrusion detection during seamless transfer, In the doctoral colloquium of Pervasive 2004, Vienna, Austria, April 18-23, 2004.
7. P. Ciancarini, Coordinating Multi-Agent Applications on the WWW: A Reference Architecture, IEEE Trans. on Software Engineering, vol.24, no.5, pp. 363-375, 2002
8. D. Garlan and D.P. Siewiorek, A. Smailagic and P. Steenkiste, Project aura: toward distraction-free pervasive computing, IEEE Pervasive Computing, vol.1, no.2, pp. 22-31, Apr-Jun, 2002
9. A.R. Tripathi, Ajanta - A System for Mobile Agent Programming, Technical Report(TR02-016), Department of Computer Science, University of Minnesota, 2002

Transformation of MPEG-4 Contents for a PDA Device

Sangwook Kim¹, Kyungdeok Kim², and Sookyong Lee¹

¹ Department of Computer Science, Kyungpook National University
Daegu, 702-701, Republic of Korea
`swkim@cs.knu.ac.kr`

² Division of Computer and Multimedia Engineering, Uiduk University
Gyeongju, 780-713, Republic of Korea
`kdkim@uu.ac.kr`

Abstract. We propose a transforming method of MPEG-4 contents in order to present the contents on PDA devices in this paper. The method reconstructs the mp4 file according to transforming a scene tree in MPEG-4 contents when the contents are authored. The method reduces the size of each object in a scene of the content for presenting it efficiently on the small interface of a PDA device, and transforms visual objects in the scene into geometry objects in order to reduce initial loading time and a size of the contents. And an original object is presented on a PDA device when the user clicks the substituted geometry object. The method was applied to a conventional authoring tool, so we could find that the method showed an efficient presentation of the contents on a PDA device.

Keywords: MPEG-4, Transformation, Authoring, Adaptation.

1 Introduction

Conventional MPEG-4 contents are generally used for a video-phone on a communication network. But, in these days, their use is very increased by growth of wireless communication. The MPEG-4 is an international standard for efficient transmission and use of multimedia data and focuses on the content-based encoding that is based on understanding of image contents. Such the content-based encoding splits image contents into object units, transmits the units, and controls and displays split respective units by a user's intention [1, 2, 3].

In MPEG-4 contents, a scene is formed by the split units that are handled individually, and a scene description language, BIFS (BInary Format for Scene), is used to describe temporal and spatial information for scene changes according to user interaction and temporal flow [1]. Users use generally intuitive MPEG-4 authoring tools because authoring contents requires professional information. Most of MPEG-4 contents authoring tools are suitable to desktop computers, and generated MPEG-4 contents from the tools are most suitable to desktop players. It is especially difficult to be presented on a PDA device for MPEG-4 contents due to a small size of a PDA screen. So conventional MPEG-4 authoring

tools need a transformation method in order to play MPEG-4 contents on a PDA device [6, 9, 10].

In this paper, we describe the method that scales MPEG-4 scene tree in authoring for presentation of MPEG-4 contents on a PDA device. The method considers a displaying screen size on a PDA device and a file size of transmitted contents to a PDA device. A scene of MPEG-4 contents can include lots of media objects (visual and aural objects). The more the MPEG-4 contents include media objects, the more a file size of MPEG-4 contents is increased. And such MPEG-4 contents are difficult to transmit and play on mobile environment. So the suggested method supports that visual objects in a scene are represented by geometry objects and an initial transmission file has only basic information about media streams. The geometry object and reducing initial file size efficiently support to play the contents on a PDA device.

This paper is organized as follows. Section 2 introduces related researches. Section 3 describes a transformation method for a PDA device. Section 4 explains implementation and comparison between scaled contents and not scaled contents. Finally, section 5 describes some concluding remarks and presents future plans.

2 Related Works

There are lots of researches on MPEG-4 contents, but researches on authoring and a transformation of contents for a PDA device are not activated yet. There are researches about authoring tools and models of contents as follows.

First, the work of design and development of MPEG-4 contents authoring system [1] provides an intuitive authoring for users on windows environments. The system supports to compose a visual and aural scene using video objects, audio objects, image objects, and text objects, and set up intuitively a scene change by a mouse click. And the system defines a scene tree for the representation of MPEG-4 contents that is consisted of object units in a scene. The scene tree is used inner data structure for authoring in the system. A scene that is represented by using a scene tree generates a scene descriptor in the form of text, and then finally MPEG-4 contents are created after the scene goes through encoding phase. The work of design and implementation of a visual MPEG-4 scene-authoring tool [4] is developed on windows environment, and the tool can store user's making scene in the data form of the server storing MPEG-4 contents. Also the authoring tool supports BIFS commands and JAVA scripts in order to user's interaction. The work of MPEG-4 authoring tool for the composition of 3D audiovisual scenes [5] generates MPEG-4 contents using 3D APIs of the OpenGL library. The tool generates 3D contents and media objects such as a box, a sphere, a cone, video, audio, etc., and then transforms the objects into the BIFS commands. Finally, the tool generates an mp4 file using the generated BIFS file and provides a preview function for a 3D MPEG-4 scene. The work of content model for mobile adaptation of multimedia information [6] proposes a new abstract model to represent and adapt multimedia to hybrid environments. The model includes a layered mapping of semantic and physical entities and is

combined under the taxonomy of multimedia adaptation to optimize end-to-end service. The adaptation taxonomy consists of two parts; semantic adaptation and physical adaptation. The semantic adaptation is based on users' and service providers' choice and it is affected by the semantic content of a presentation. The physical adaptation is based on physical QoS and the characteristics of media objects consisting multimedia contents. The work of adapting multimedia internet content for universal access [7] presents a system adapting multi-media web documents to optimally match the capabilities of the client device requesting it. The system shows that multimedia contents are adapted by using two components; a representation scheme that provides a multimodal and multi-resolution representation hierarchy for multimedia and a customizer that selects the best content representation to meet the client capabilities while delivering the most value. The scalability for adaptive MPEG-4 contents describes an adapting technique for MPEG-4 contents on various service environments. The adapting technique includes adaptations of each media object in a scene and streaming service using network bandwidth. But the technique doesn't support to author MPEG-4 contents with interactive capability for mobile devices such as a PDA device, and it introduce capability which MPEG-4 contents is reconstructed using media objects in a scene.

Most of above-described researches are difficult to suitably support contents authoring for mobile devices. Conventional MPEG-4 contents need an adaptation in order to be used on mobile devices. So, in this paper, we propose a transformation method that uses characteristics of a mobile device in authoring contents.

3 Transformation of MPEG-4 Contents

Conventional MPEG-4 contents are difficult to present on a PDA device because they are optimized at desktop computers. Specially, they have a problem on a displayer due to difference on a screen size of a PDA device and conventional devices. So, in this paper, we consider two sides on authoring MPEG-4 contents. First, we consider reduction of a physical size of each object in each scene because of a small screen size of a PDA device. Second, we consider a reduction of transmission quantity on content's stream for fast initial presentation on a PDA device.

The figure 1 presents an authoring course of MPEG-4 contents for a PDA device. The authoring course includes 5 steps as following; arrangement of media objects, and setting up of object's attributes, management of scene description information, and scaling objects, adjustment of scene tree, encoding of BIFS, and generation of an mp4 file. The first step (arrangement of media objects, and setting up of an object's attributes) supports to compose a scene using geometry objects which represent media objects. The object's attributes which set up visual characteristics at each media object are established by a user's a mouse click or simply textual input on a dialog box. The second step (management of scene description information, and transforming objects) supports to generate a scene

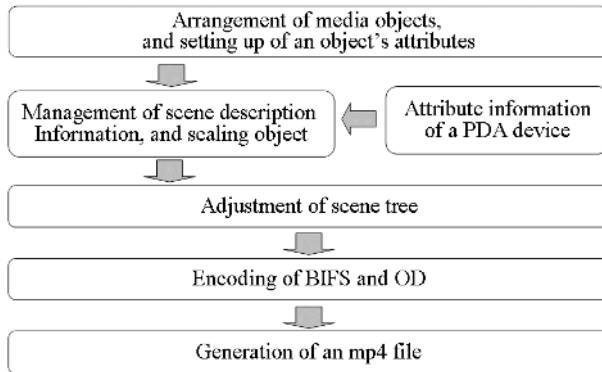


Fig. 1. An MPEG-4 authoring course for a PDA device

tree and scale a media object. The generated scene tree is used to manage scene description information, and each node in the scene tree has information on a temporal and spatial relation among objects as following; a geometry object, an image object, and a video or audio object. Whenever an object is generated in an authoring space, an object node is added in the scene tree. And an attribute object is added as a subordinate node of the object node in the scene tree on setting up temporal and spatial information. The subordinate node has information about form, location, playback time, etc. of a media object. Such information of object's attributes is represented as an object descriptor(OD) in MPEG-4 contents. The third step (adjustment of a scene tree) supports to modify information of a node in the scene tree and physical size of a media object in order to generate MPEG-4 contents for a PDA device. The fourth step (encoding of BIFS and OD) supports to generate BIFS and OD contents from information of each node in the adjusted scene tree using rules generating BIFS and OD. The BIFS is represented as a textual form, and then they are encoded as a binary form. The fifth step (generation of an mp4 file) supports to generate an mp4 file from multi-flexing an encoded binary data (BIFS and OD).

The figure 2 presents a change of a scene tree using the suggested transforming technique. The upper part and lower part in the figure presents an original scene tree and a scaled scene tree respectively. In the figure, the part showed dotted line presents an image or video object, and the part is transformed into a geometry object by the suggested transforming technique. In the figure 2, the GO1 object or GO2 object means a visual object as followings; rectangle, triangle, line, etc. In a scaled scene tree, the geometry object has subordinate nodes that have information in order to activate a corresponding media object in an original scene tree. So, when contents are played, an original media object described in an original scene tree is replayed if user clicks the generated geometry object in the scaled scene.

In order to generate contents for a PDA device, information of media objects to be scaled is extracted from a scene tree. The information is made as a form of

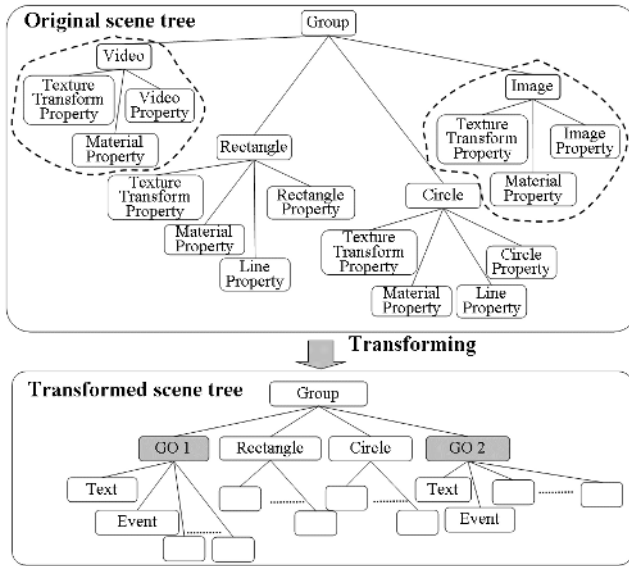


Fig. 2. Change of a scene tree by the transforming technique (A rounded rectangular in a tree presents a node on a scene tree)

linked list, and is used to scale and transform a media object. The transforming a media object transforms the media object into a geometry object, and generates an event which supports to play an original media object instead of showing the geometry object. The event is added as a subordinate node of a corresponding node in a scaled scene tree.

The figure 3 presents a transforming process of a scene tree. In this process, a screen size for playback is decided by using location values of media objects in each scene of contents. The minimum screen size for playback on a PDA device is computed by leftmost, rightmost, uppermost, and lowermost location of each media object in a scene. Also, the size is used to compute a reduction rate in contrast with a screen size for playback of original contents. The reduction rate is applied to reduce an external size of each geometry object (line, circle, rectangular, etc.) in a scene, and adjusts a coordinates' value of the geometry object, too. The rest of a text, an image, an audio, a video, and an event object are scaled as followings; In the case of a text object, a big reduction can make user not seeing itself on a PDA device. So, in order to avoid this problem, a reduction rate for exceeding a certain font size should be considered. In contrary of a text object, an image object dose not reduced, but if a size of an image object exceeds a screen size of a PDA device, a corresponding node in a scene tree is removed in the transforming technique. So, when an external size of an image object is simply smaller than a screen size of a PDA device, the image object is play backed. In the case of a video object, a size of the object (QCIF(144x176) or CIF(288x176)) is mostly smaller than a screen size of a PDA device. However,

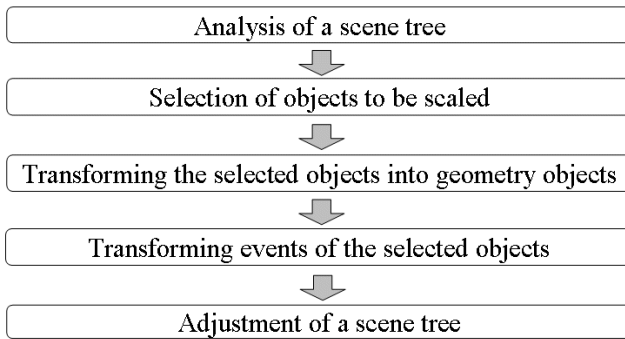


Fig. 3. Transforming process of a scene tree

lacking resources on a media player in a PDA device can bring about quality deterioration on playback. This problem is solved by encoding scalability, but is not considered in this paper. In the case of an image or a video object with temporal information, if the objects should be played back at same location on a PDA screen in regular sequence, a firstly played object is substituted for a geometry object. The geometry object indicates a text 'I'(image) or 'V'(video) according to a really played object. Activation of the geometry object by a users' click ignites playback of connected media objects with time.

4 Implementation and Results

In this paper, the proposed transforming method is applied to a conventional authoring tool [8] for presentation of MPEG-4 contents on a PDA device (screen size: 288x352). Authored MPEG-4 contents using the proposed transforming method indicate geometry objects instead of visual objects on a PDA device. The geometry objects ignite playback of a connected real media object by users' click.

The figure 4 presents MPEG-4 contents on a PC and a PDA respectively. In the figure, left side figure present a complete scene of MPEG-4 contents on a PC, and right side figure present a partial scene of MPEG-4 contents on a PDA due to the suggested transforming technique. However, we could find that the method showed an efficient presentation of the MPEG-4 contents on a PDA device.

The figure 5 presents a comparison between pre-application and post-application of the transforming method in authoring MPEG-4 contents. In the experiment, we compared a file size of an mp4 and BIFS according to increase in number of an image object or a geometry object; in the figure 4, the GO is shorts for a geometry object, and the IO is shorts for an image object. The experiment showed that a file size of BIFS is not enough of a difference whether a geometry object is included or an image object is included in contents. However, a size



Fig. 4. Presentation of MPEG-4 contents on a PC and a PDA

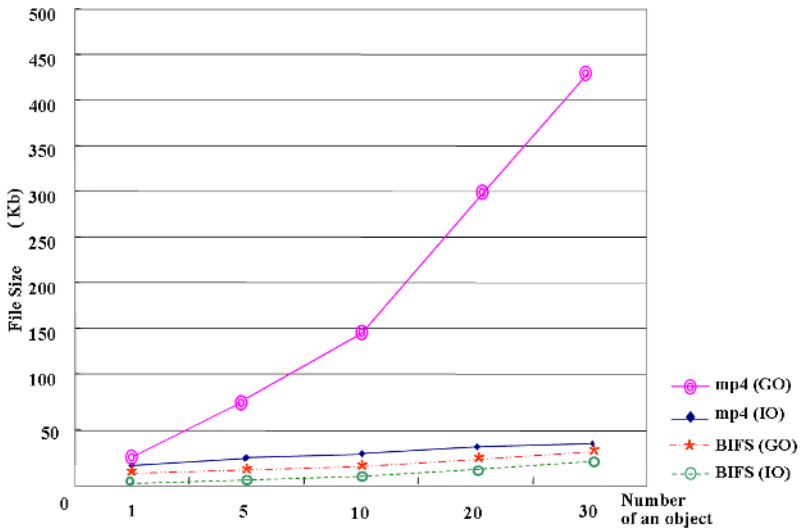


Fig. 5. Comparison between pre-application and post-application of the transforming method (GO: geometry object, IO: image object)

of an mp4 file has all the difference between using an image object and using a geometry object according to increase in number of an object.

In view of the result, scaled contents have a very small file size due to not transmitting a media object. Namely, only BIFS of scaled contents is enough of construction for whole scene and efficient playback on a PDA device. Generally, MPEG-4 contents with media objects are multiplexed with BIFS, OD, and media objects for generation of an mp4 file, which are transmitted to a player device. So the more contents have media objects, the more an mp4 file size is increased. But, a scene of scaled MPEG-4 contents uses geometry objects instead of visual objects; a PDA device receives only BIFS information for initial scene generation, and a user can play back an original visual object according to a user’s click.

Such BIFS has not a large size due to not including a visual object, and it can reduce usage on resources of a PDA device and network bandwidth.

5 Conclusions

In this paper, we proposed a transforming method in order to present MPEG-4 contents on a PDA device. The transforming method selects visual objects to be scaled in a scene tree and transforms the selected objects into geometry objects or smaller objects to fit on a PDA screen for automatic generation of an mp4 file. Also the transforming method has hardly information lose of original contents. It is easily applicable to a conventional authoring tool, supports user's authoring contents with user not having information on BIFS, and supports to directly author contents for a PDA device on desktop environment.

Future works are development of an adjustment technique on a scene with events and an encoding technique for scalability on visual objects.

Acknowledgement. This work was supported by Korea Research Foundation Grant (KRF-2003-002-D00304).

References

1. Cha, K., Kim, H., Kim, S.: The Design and Development of MPEG-4 Contents Authoring System. *J. of Korea Information Science Society*, Vol. 7. (2001) 309–311
2. WG11(MPEG), MPEG-4 Overview (V.16 La Baule Version) document. ISO/IEC JTC1/SC29/WG11 N3747 (2000)
3. WG11(MPEG), MPEG-4 Overview (V.18 Singapore Version) document. ISO/IEC JTC1/SC29/WG11 N4030 (2001)
4. Shieh, M., Perngand, K., Chen, W.: The Design and Implementation of A Visual MPEG-4 Scene-Authoring Tool. *Proc. of Workshop and Exhibition on MPEG-4* (2001) 21–24
5. Daras, P., Kompatsiaris, I., Strintzis, M.: MPEG-4 Authoring Tool for The Composition of 3D Audiovisual Scenes. *Proc. of Second Int. Workshop on Digital and Computational Video* (2001) 110–117
6. Metso, M., Koivisto, A., Sauvola, J.: Content Model for Mobile Adaptation of Multimedia Information. *Proc. of IEEE 3rd Workshop on Multimedia Signal*. (1999) 39–44
7. Mohan, R., Smith, J., Li, C.: Adapting Multimedia Internet Content for Universal Access. *IEEE Transactions on Multimedia*, Vol. 1. (1999) 104–114
8. Cha, K., Kim, S.: MPEG-4 STUDIO: An Object-Based Authoring System for MPEG-4 Contents. *Multimedia Tools and Applications*. (2003)
9. Lee, S., Cha, K., Kim, S.: An MPEG-4 Contents Authoring for Mobile Devices. *Proc. of 2003 HCI* (2003) 402–405
10. Cha, K.: A Scalability for Adaptive MPEG-4 Contents. *Dissertation for The Degree of Doctor of Philosophy*, Kyungpook National Univ., Korea (2003)

Semantic Retrieval in a Large-Scale Video Database by Using Both Image and Text Feature

Chuan Yu¹, Hiroshi Mo², Norio Katayama²,
Shin'ichi Satoh², and Shoichiro Asano^{1,2}

¹ University Of Tokyo

² National Institute of Informatics

{wusen,mo,katayama,satoh,asano}@nii.ac.jp

Abstract. The paper demonstrates a new retrieval method of intergrating both image features and text informations derived from video data. We compare not only image similiarity, also narrow the retrieval sets in advance by employing searching in text keywords. Users inputs also performs the key role in improving retrieval accuracy.

1 Introduction

So far robust content-based image retrieval in the large-scale video/image database is still a challenging problem. Many of developed techniques operate under specific constraints about the type of image data they operate upon. Frequently, such constrains have been put on the retrieval of images containing certain objects. Early work used user-defined sketches for retrieval [1], whereas some approaches perform content-based retrieval of images using object models [6]. [3] presented the way between low and high level features to classify images into different categories according to regions. Regardless of manmade structures which [2] used,[7] showed the method to segment image or video data using spatial information. There is no clear consensus about which technique to use for a general image retrieval system. The answer to this problem depends on many factors, such as the number and complexity of objects present in an image, the amount of a priori information about the scene, and the purpose the retrieval system is. In this paper, we explore how user involvement with an image retrieval system using various features of this system can improve performance. As a fundamental job of our proposed news video digest system that gives the ability to retrieve correct parts of news video from a large-scale video database to where user shows the interest, we investigate the effect of feature integration, the keywords in news video, and the assistant of user input. As mentioned above, features extracted from different techniques emphasize image attributes in different domains.

First of all, we calculate different features of image for boosting retrieval performance in advance. Further, the keyword input plays a great role on image and shots extracting with relative topic. Then the user input indicates which object on the certain image is most interested, and the combination of keywords and

features retrieval will be performed to create the search result. Most of current view-based retrieval techniques only analyze image data at a lower level on a quantitative basis for color and/or texture features. These techniques are geared towards retrieval on overall image similarity, especially for images containing natural objects. There is no clear consensus which technique to use for a general image retrieval system yet, the answer to this problem depends on many factors, such as image object, the type of image, and the most of all, the purpose of research. Our research gives a new answer to solve the problem by integrating image feature and text information. The rest of paper is organized as follows. Section 2 details the structure of system. Section 3 describes the detail of images features and keyword and how to integrate them to retrieve image. Section4 shows the search result respectively. The last section outlines the method and the future work

2 Structure of System

The system work flow shows in the Fig.1. We first segment news video into video shots and select representative frame for each video, in the meantime, the video keywords for each shots are captured too. Then image features of all image frames are calculated in advance offline before retrieval. Thus the original image database including image, keywords, and features is created.

The retrieval interface accepts the individual or multiple keywords inputted by user and sends the query to image database and returns the retrieved result back to user. User can also choose any interest image randomly selected by interface and retrieve global similar images through single or integrated image

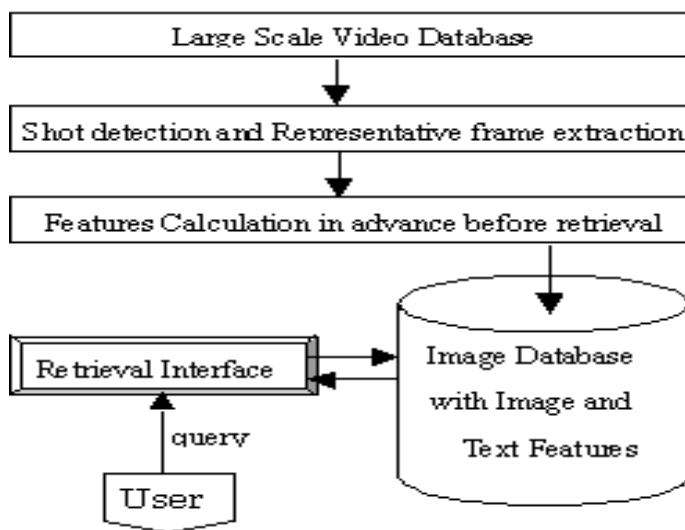


Fig. 1. System work flow

features. In addition, the effect of supplying more than one query image is considered. On the condition that user choose more than one image to perform more accuracy retrieval, the distance between image results and any selected image is computed and the result of image cluster is reconstructed.

3 Individual and Integrated Features

We employ two kinds of color histogram and edge histogram features to calculate global feature composition of an image.

3.1 Image Features

Color Histogram: The traditional color histogram is very useful because it is invariant to translation and rotation of the images and normalizing the histogram leads to scale invariance. Here we use compute color histogram features in two different ways:

1. The image is divided into 2×2 sub-images evenly, and the histogram of RGB of each sub-image is calculated on the axis of one dimensional RGB space that is divided into 16 partitions respectively. Thus there are $16 \times 3 \times 4 = 192$ bins. The color histogram feature is designated as ω_1 .
2. The image is divided into 2×2 sub-images too, and the joint histogram of every sub-image is calculated in the three dimensional RGB spaces which is divided into $4 \times 4 \times 4 = 64$ sub-cube space. Thus there are $64 \times 4 = 256$ bins. The cube histogram feature is designated as ω_2 .

Edge Histogram: Edge in the image is considered an important feature to represent the content of the image. Human eyes are known to be sensitive to edge features for image perception. The normative part of the edge histogram descriptor consists of 80 local edge histogram bins [4],[5].

There are five different types of edge in the image, four directional edges and a non-directional edge. Four directional edges include vertical, horizontal, 45 degree, and 135 degree diagonal edges (1)→(5). To detect each of them, we can use five edge filters corrsponing to different edge.

$$\textit{Vertical} - \textit{filter} = \begin{vmatrix} 1 & -1 \\ 1 & -1 \end{vmatrix} \quad (1)$$

$$\textit{Horizontal} - \textit{filter} = \begin{vmatrix} 1 & 1 \\ -1 & -1 \end{vmatrix} \quad (2)$$

$$\textit{Diagonal}45^\circ - \textit{filter} = \begin{vmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{2} \end{vmatrix} \quad (3)$$

$$\textit{Diagonal}135^\circ - \textit{filter} = \begin{vmatrix} 0 & \sqrt{2} \\ -\sqrt{2} & 0 \end{vmatrix} \quad (4)$$

$$\textit{Noneedge} - \textit{filter} = \begin{vmatrix} 2 & -2 \\ -2 & 2 \end{vmatrix} \quad (5)$$

Usually, the image is splited into 4×4 sub-images, and every sub-image is divided the sub-image into a fixed number of image-blocks. That is, the size of the image-block is proportional to the size of original image to deal with the images with different resolutions. Equations (6) and (7) are used to divide sub-images.

$$x = \sqrt{\frac{\text{heightofsubimage} \times \text{widthofsubimage}}{\text{desirednumber}}} \tag{6}$$

$$\text{block} - \text{size} = \lfloor \frac{x}{2} \rfloor \times 2 \tag{7}$$

Here, the image-block is further divided into four sub-blocks. Then, the luminance mean values for the four sub-blocks are used for the edge detection. The next convolution equation (8) is used to extract the number of bin existing in the each sub-image.

$$\text{edge}(i, j) = \left| \sum_{k=0}^3 L_k \times \text{edge} - \text{filter} \right| \tag{8}$$

(i,j) stands for the number of image block, L_k is the average of luminance of image block. Using different edge filter, the edge bin can be computed. Since there are 16 sub images and 5 types of edge, therefore the edge histogram has 80 bins. The edge histogram feature is designated as ω_3 .

3.2 Keywords

The keywords are related to every shots of news video data. Those keywords are not a absolute copy of announcers voice in news video. We create the keywords database by choosing important words. A typical keywords relating to a shot looks like:

68740;keywords1:1;keywords2:1,;keywords3:1,;keywords4:1,;

The keyword composes of time, keywords and the frequency of keywords appear. The query of keyword input by user can generate relative image cluster about certain topic, which can be people, places, or sports and political affairs.

3.3 Integrating Features

The keyword composes of time, keywords and the frequency of keywords appear. The ability to extract and describe distinct objects in a complex scene is crucial for image understanding, in particularly it is difficult when deal with a large scale image database captured from the very complex news video data.

The color histogram is suitable for global recognition of image, and the edge histogram is suitable for structure analysis. The aim is to explore if feature integration using (i) color histogram (ii) edge histogram results in better performance than using those features individually. We employ simpler feedback

that users get involved in the retrieval process too to choose one or more query images to improve retrieval accuracy.

First of all, Interface returns search results according to the inputted single or multiple keywords. Then user can designate one or more image queries to reconstruct the new result dataset by using integrated image features. The simplest way is user input keyword as search query to retrieve image. User can input more than one keyword to retrieve image, k_i is the result retrieved by i th keyword. α is the constant coefficient to control whether the features is to be integrated.

$$W = \prod_{i=0}^m k_i \alpha_i \quad (\alpha_i = 0, 1) \quad (9)$$

We use a linear combination (10) of features to complete retrieval via integration of features. W stands for the final retrieved result. ω_i is the results retrieved by using traditional color histogram, cube histogram or edge histogram respectively, α has the same meaning as above to control how many text queries are involved. We get different result sets through selecting different image and/or keywords parameters. The integration of features is not only among image features or text query respectively, but also can be expand to among image features and text queries (11).

$$W = \sum_{i=1}^n \omega_i \alpha_i \quad (\alpha_i = 0, 1) \quad (10)$$

$$W = \sum_{i=1}^n \omega_i \alpha_i \cap \prod_{j=1}^m k_j \beta_j \quad (\alpha_i, \beta_j = 0, 1) \quad (11)$$

Since our text database is not absolute one on one map, it means sometimes the retrieved image does not include the object designated by query, in the case, the user assistance is necessary and valuable. User can specify more than one image and retrieve either in individual way or in integrating way. System retrieves dataset firstly by text retrieval, and then compute distance between every image in the dataset and the every query image by using individual or integrated image feature retrieval. Thereafter, the final dataset is created by the completion of iteration of computation of every query image.

4 Experiment and Discussion

Our image database composes of 45,000 images, keywords and features data calculated in advance. The size of image is 352×240 . The images include all representative frames of 30 minutes video data of news 7 of NHK of half year from 10,2003 to 3,2004.

The retrieval speed is relatively high since all the images features are calculated in advance. In generally, it takes around 8 seconds to retrieve similar images from 45,000 images using any of images features, and it takes about 2 seconds to retrieve by one keyword input and 4 seconds by 2 keywords input and 6 seconds by 3 keywords. The total calculated time shows in the Table 1.

Table 1. Features computing time

Monthly Features	2003/10	2003/11	2003/12	2004/01	2004/02	2004/03
Computing Time(s)	603	569	631	660	465	706



Fig. 2. Multi keywords retrieval result



Fig. 3. Retrieval results in 2 keywords

The fig. 2 shows the result in 2 keyword inputs. There are 69 results corresponding to keywords “Koizumi” and “Iraq”. And there would be 282 results if we only inputs “Koizumi”.

The fig. 3 shows the retrieval result in joint keywords “Matsui”, “Big League”. User maybe not be satisfied with the result since there are too many irrelevant results.



Fig. 4. Integrating all 3 image features

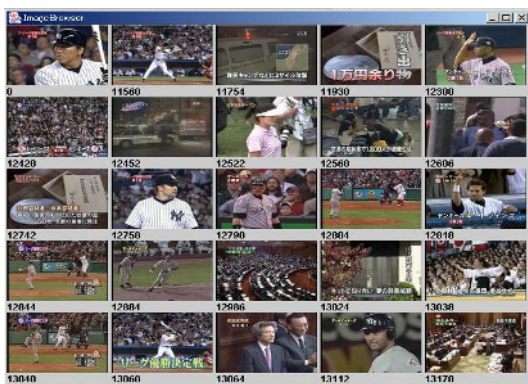


Fig. 5. Integrating 2 image features retrieval

The fig. 4 and fig. 5 shows the result of integrating image and text features. In the fig. 4, user select top left image as query to retrieve similar image by integrating all 3 types of image features. And fig. 5 shows the result user select top left 2 images as query to retrieve similar image by only using cube color histogram. The latter has the better performance than the former, we can point out that it is not for sure integrating more feature means higher accuracy. To different kind of objects or proposes, we should select different strategy of integrating features. But integrating image and text features with the assistance of user input can obviously improve the image retrieval performance, and it makes us closer to do content-based retrieval. If we add object extraction in the image to let user has the capacity to feedback the object in the image, then the retrieval process will be more truly semantic and higher accuracy.

5 Summary and Future Work

The difficulty of image and video retrieval arises from a number of issues, such as the complex mix of manmade and natural objects in an image. We studied the features and how to integrate them to improve performance, the edge histogram are suitable for object detection and histogram is easy to use in global comparison. It was observed that feature integration enhanced retrieval accuracy in general. And the user assistant also proved valuable. The method provides a useful tool to achieve content-based image retrieval. The goal is not only to retrieve similar image, underlying this work, we will add precise object extraction on the system, and even to analyze video shot itself. We want not only to achieve image retrieval, but also content-based video retrieval. The work is still in process to the goal of on-demand image and video retrieval system.

References

1. A.D. Bimbo, P. Pala and S. Santini, "Visual Image Retrieval by Elastic Deformation of Object Sketches," in IEEE Symposium on Visual Languages, St. Louis, MO, pp.216-223,1994
2. Q. Iqbal and J.K. Aggarwal, "Retrieval by Classification of Images Containing Large Manmade objects using perceptual grouping," Pattern Recognition, vol.35, pp.1463-1479, July 2002
3. J.H. Lim and J.S. Jin, "Semantic Discovery for Image Indexing," ECCV 2004, LNCS 3021,pp 270-281, 2004
4. A.K. Jain and A. Vailaya, "Image retrieval using color and shape," Pattern Recognition, Vol. 29, No. 8, pp.1233-1244, 1996.
5. S.J. Park, D.K. Park, C.S. Won, "Core experiments on MPEG-7 edge histogram descriptor," MPEG document M5984, Geneva, May, 2000.
6. J. Sivic, F. Schaffalitzky and A. Zisserman, "Object Level Grouping for Video Shots," ECCV 2004, LNCS 3022, pp.85-98, 2004
7. J. Wang, B. Thiesson, Y.Q. Xu and M. Cohen, "Image and Video Segmentation by Anisotropic Kernel Mean Shift", ECCV 2004, LNCS 3022, pp.238-249,2004

Performance of Correlation-Based Stereo Algorithm with Respect to the Change of the Window Size

Dong-Min Woo¹, Howard Schultz², Edward Riseman², and Allen Hanson²

¹ Myongji University, Yongin, 449-728, Korea
dmwoo@mju.ac.kr

² University of Massachusetts, Amherst, MA 01003, USA
{hschultz, riseman, hanson}@cs.umass.edu

Abstract. Correlation-based stereo matching is very important to the generation of 3D terrain model. One of the difficulties in this stereo matching is the selection of the window size, since there are two competing factors that must be balanced in any stereo reconstruction process - perspective distortion and stereo matching error. This paper presents how the correlation window size affects the accuracy of stereo matching and suggests a tool to tune the window size in a given image domain. To facilitate the analysis proposed in this paper, we use photo-realistic simulation methodology to generate a pair of photo-realistic synthetic images of the terrain from a pre-acquired DEM(Digital Elevation Map) and ortho-image, which can be served as the pseudo ground truth. We performed 3D reconstruction on synthetic images of a natural terrain with Terrest system and carried out the evaluation of the correlation window on DEM accuracy. Experimental results are consistent with our strong expectation about two competing factors and show that our approach can be a useful tool to tune the window size.

Keywords: Stereo matching, 3D terrain model, DEM.

1 Introduction

There has been a significant body of research in 3D reconstruction using stereo image analysis and a variety of algorithms have been developed for diverse range of applications that include satellite and aerial remote sensing, mobile robotics, industrial inspection and object modeling [1,2,3,4,5,6]. Correlation-based stereo matching [6] is particularly useful to generate 3D terrain model, since a dense array of depth values can be calculated over a region.

In spite of significant advances in stereo image analysis, there still remain serious unsolved problems and sources of errors in the stereo matching process, such as photometric variation, occlusion, repetitive texture and lack of texture in the correlation window [7]. These problems cause matching error, which make an abrupt peak called "spike" in disparity map. Another type of error comes from perspective distortion. Although the centers of two correlation windows

are aligned, the other pixels in the windows may not be aligned. This pixel misalignment is shown more significantly for the pixel farther from the center [8]. The pixel alignment does not make serious error like "spike", but it still causes the type of error, which is shown evenly across the entire disparity map.

In computing a disparity map by using correlation-based stereo matching, a dilemma occurs when the size of the correlation window is selected. To increase robustness to matching error, it is usually noted that the correlation window should be as large as possible. To minimize the effects of perspective distortion, the window should be as small as possible [9]. To support the selection of the window size in this tradeoff situation, this paper presents how the correlation window size affects the accuracy of stereo matching and suggests a tool to tune the window size in a given image domain. To facilitate the analysis proposed in this paper, we use photo-realistic simulation [10] to generate photo-realistic synthetic images through pseudo ground truth, rather than using ground truth of disparity map [11]. With this approach, we conduct a quantitative accuracy analysis, more focused on 3D reconstruction of real terrain.

3D reconstruction is performed on synthetic images with Terrest system [12], which is a correlation-based 3D reconstruction system developed at University of Massachusetts. We employed multi-resolution scheme [5], referred to as hierarchical, or pyramid processing, to our Terrest system. This reduces the chance of encountering false matches in addition to saving computation time. We also used subpixel interpolation, which enables us to estimate accurate disparities.

2 Terrest System

2.1 Normalized Cross-Correlation

The goal of image matching is to find a disparity map $D_i(i, j)$ that maps the pixels in the epipolar resampled reference image $I_R(i, j)$ into the epipolar resampled warped image $I_W(i, j)$ such that each pixel pair sees the same spot on the object, i.e., $I_R(i, j)$ and $I_W(i + D_i(i, j), j)$ view the same spot on the surface.

One of the most robust and commonly used match score is the cross-correlation coefficient $p(i, j, d_i, d_j)$ between a rectangular window centered at $(i, j)_R$ and a similar window centered at $(i + d_i, j + d_j)_W$. By definition, $p(i, j, d_i, d_j)$ is given by

$$p(i, j, d_i, d_j) = \frac{Cov[I_R(i, j), I_W(i + d_i, j + d_j)]}{\sqrt{Var[I_R(i, j)]Var[I_W(i + d_i, j + d_j)]}}, \quad (1)$$

where $I_R(i, j)$ and $I_W(i + d_i, j + d_j)$ are the pixel intensities, $Cov[I_R(i, j), I_W(i + d_i, j + d_j)]$ is the covariance between the windows, and $Var[I_R(i, j)]$ and $Var[I_W(i + d_i, j + d_j)]$ are the variances within each window. In this formula, $p(i, j, d_i, d_j)$ does not depend on the positions of the pixels within the window, and all pixels contribute equally to the match score. When the masks are aligned, pixels near the center of the mask are less affected by perspective distortions.

2.2 Subpixel Interpolation

Reconstruction accuracy depends directly on the disparity map accuracy; therefore, significant improvements can be achieved by computing disparities to subpixel accuracy. Subpixel registration schemes rely on the assumption that near the true disparity d_t , the computed match scores are estimates of a smooth function $f(d)$ and that d_t satisfies the condition,

$$\left. \frac{df(d)}{dd} \right|_{d=d_t} = 0. \quad (2)$$

Thus, an estimate of the optimal disparity d_* is found by approximating $f(d)$ with a model $f(d; c_0, c_1, \dots)$, solving for the coefficients c_0, c_1, \dots , and setting d_* to the value of d that optimized the model,

$$\left. \frac{df(d; c_0, c_1, \dots)}{dd} \right|_{d=d_*} = 0. \quad (3)$$

Practically, a parabolic model or a Gaussian model can be used as a smooth function f .

2.3 Multi-resolution Scheme

For terrain it is generally true that large objects have large disparities and small objects have small disparities. When the resolution of the I_R and I_W images are reduced, smaller features disappear. Thus, only small scale disparities are lost when the low resolution images are correlated. Once the large scale disparities are recovered, the small scale disparities are recovered by processing the high resolution images and restricting the disparity search to perturbations about the previously recovered disparities. This refinement process results in a significant reduction in the amount of computation, which also reduces the chance of encountering false matches in addition to saving time. This sequential processing from low to high resolution image pairs is referred to as hierarchical, or pyramid processing.

An image pyramid is a set of images $I^{(0)}, I^{(1)}, \dots$ of progressively diminishing resolution that are derived from a common parent image I . Resolution reduction is accomplished by smoothing the previous layer and then selecting every other pixel. Usually we use 4 level pyramids and the images are reduced by convolving 3x3 Gaussian kernel and selecting every other pixel.

3 Photo-Realistic Simulation Analysis

3D reconstruction described in this paper produces a dense array of elevation estimates, which typically involves millions of samples. Consequently, a comprehensive evaluation of correlation-based stereo matching result requires an equally dense array of ground truth. A few manually gathered ground control

points will not provide a sufficient number of samples to test the validity of the 3D reconstruction result.

To provide a detailed analysis of correlation-based stereo matching, we create a pseudo ground truth data set. The process begins with an existing DEM and ortho-image. A photo-realistic ray-tracing program is used to synthesize images of the surface from arbitrary viewpoints. Next, the synthetic images are applied to correlation-based 3D reconstruction system and the result DEM is regenerated. The regenerated DEM can then be compared on a point-by-point basis to the pseudo ground truth. With this approach, we can quantitatively determine the accuracy of the regenerated DEM and then evaluate the performance of correlation-based stereo matching process.

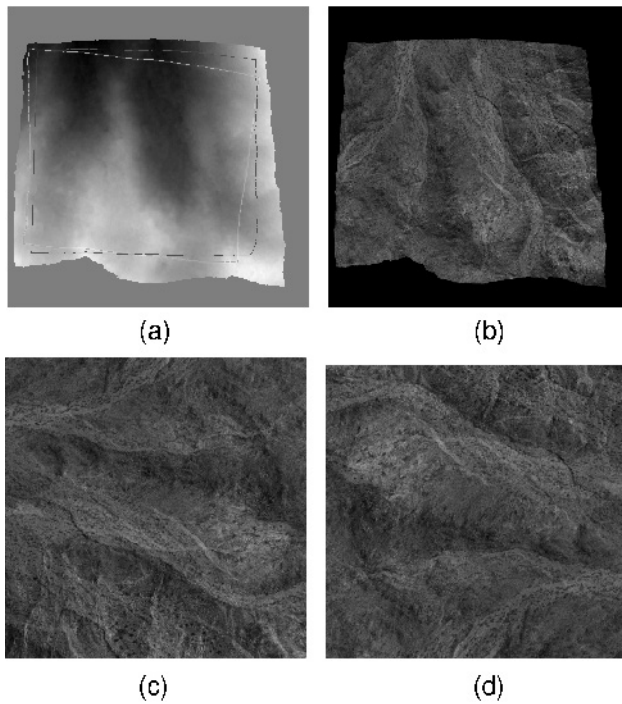


Fig. 1. Generation of synthetic images using pseudo ground truth: (a) DEM (b) ortho-image (c) synthetic image 1 (d) synthetic image 2

For this study, a small region covering area of 610m x 610m was selected for the pseudo ground truth. To remove another complicating factor, we select an area of terrain with less occlusion. Otherwise there might be terrain contexts that would actually deviate from expectation, but only due to unmodeled characteristics. Fig. 1 (a) and (b) depict the pseudo ground truth DEM(Digital

Elevation Map) and ortho-image, originally generated by aerial images of desert scene. The contour overlaid in Fig. 1 (a) is the projected area for the two synthetic images, as shown in Fig. 1 (c) and (d). Our synthetic images are generated as nadirs views which are very common in aerial image acquisition.

4 Experimental Results

Based upon very clear theoretical principles we have strong expectations for the quantitative results of this experiment in parametric optimization in DEM accuracy with respect to the size of the correlation window [9]. There are two competing factors that must be balanced in any stereo reconstruction process. The error due to perspective distortion becomes larger as the size of the window grows. Pixels in the two correlation windows are increasingly misaligned, even when their center pixels are perfectly aligned, due to larger alignment errors of pixels that are distant from the window center [8]. To take a concrete example, consider a planar surface imaged at oblique angles inward, where opposite sides of the surface are foreshortened in the two correlation windows. Therefore, perspective distortion increases and match accuracy decreases with window size. This type of DEM error will appear evenly across the entire DEM. This implies that to minimize this effect of perspective distortion, the size of the window should be reduced, but not without limit. A competing factor is that as the size of the window becomes smaller, we are increasingly likely to generate matching errors (for a variety of reasons), each of which potentially can have very large spike error as an outlier.

In carrying out this evaluation of the correlation window on DEM accuracy with Terrest, the window size can be applied differently at each level of the multi-resolution matching scheme, which significantly complicates the analysis. To simplify this experiment, we use the same window size at each level.

Fig. 2 shows DEMs generated by 3D reconstruction as the correlation window is varied. Spike outliers are removed in a standard manner through median filtering of local neighborhoods. With a smaller window size, the spike types of outliers are observed far more frequently. These outliers can be detected by median filtering of the neighboring disparities and eliminated by substitution of the interpolated disparity. Fig. 2 shows DEMs generated by 3D reconstruction including spike elimination. With a 15 x 15 window the DEM is shown to be very smooth.

Table 1 summarizes the quantitative assessment of accuracy. Average error is associated with and dominated by contributions across all pixels, while the rms error is dominated by spike outlier errors. The average DEM error with spikes does not change significantly across small window sizes (beyond 3 x 3) because the perspective distortion increases, but the matching error decreases with modest increases in window size. As for the rms error, as expected we have the largest with 3 x 3 window and the smallest with 13x13. It appears that the best overall choice is a 13 x 13 window size without a spike removal process. When median filtering is used for spike removal, the DEM error - both

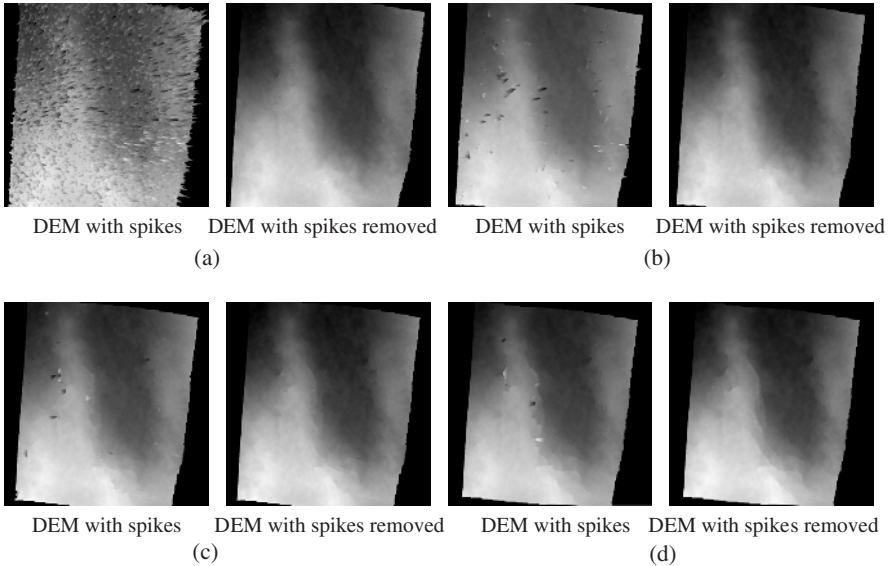


Fig. 2. DEM generated by using varying size of correlation windows in 3D reconstruction with Terrest: ((a) 3×3 (b) 7×7 (c) 11×11 (d) 15×15)

average and rms - monotonically increases with increasing window size from 5×5 onward. In this case the error resulting from perspective distortion apparently dominates. One exception is with 3×3 window, where we get larger average DEM error than with 5×5 window. This occurs due to the many outlier spikes, as shown in Fig. 2 (a), and even the successful substitution of many interpolated values the performance of 3D reconstruction is still degraded. Obviously, 5×5 is the best choice for the window size when the median filtering is employed. This result of our analysis is absolutely consistent with our rough qualitative expectation, but the parameter is quantitatively optimized using our photo-realistic simulation tool. There would have been no way for us to have determined that a 5×5 window was best, and that average error would increase by approximately 20 percent and 60 percent for 7×7 and 9×9 windows, respectively. Thus, while other correlation window sizes might be reasonable for different types of terrain or imaging conditions, we are able to tune our algorithm for this image domain.

5 Conclusions

In this paper, we have evaluated effects of correlation window size on the accuracy of correlation-based stereo matching algorithm. In fact, the fundamental tradeoff in stereo matching is the selection of window size. Decreasing the window size decreases pixel alignment errors and increases errors caused by stereo

Table 1. DEM error as correlation window size is varied. Error in reconstruction measured against ground truth (unit: meter)

window size	<i>Error of DEM with spike</i>		<i>Error of DEM without spike</i>	
	Average error	Rms error $\sqrt{E [e^2]}$	Average error	Rms error $\sqrt{E [e^2]}$
3 x 3	3.820631	6.399680	0.161612	0.386285
5 x 5	0.785736	2.874122	0.125846	0.253184
7 x 7	0.387964	1.806250	0.154254	0.271174
9 x 9	0.337733	1.311774	0.204980	0.365548
11 x 11	0.340495	1.060019	0.260392	0.452575
13 x 13	0.364696	0.863924	0.315674	0.535606
15 x 15	0.428795	0.975971	0.381458	0.646191
17 x 17	0.491957	0.964200	0.463993	0.797337
19 x 19	0.560241	1.043378	0.540334	0.931272

matching. In this context, our work can be a valuable tool to determine one of the most critical parameters of stereo reconstruction, the size of the correlation window, so that Terrest was customized to the particular quality of the terrain being processed. To provide a detailed analysis of correlation-based stereo matching, we use photo-realistic simulation methodology to generate photo-realistic synthetic images through pseudo ground truth. With this approach, we can conduct a quantitative accuracy analysis of 3D reconstruction result.

In this paper, a natural terrain is selected for the experimentation. However, an urban area is an interesting scene to be further analyzed, since 3D reconstruction of buildings presents many challenges due to discontinuities at their boundaries. Proper selection of the window size can significantly reduce the reconstruction error of buildings and we conclude that there may be a very practical need for tuning the size of the correlation window in stereo matching of urban images.

References

1. Kanade, T., Okutomi, M.: A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 16 (1994) 920-932
2. Ayache, N., Faverjon, B.: Efficient Registration of Stereo Images by Matching Graph Description of Edge Segments. *Int'l J. Computer Vision* (1987) 107-131
3. Agouris, P., Schenk, T.: Automates Aerotriangulation Using Multiple Image Multipoint Matching. *Photogrammetric Engineering and Remote Sensing*, Vol. LXII (1996) 703-710
4. Fua, P., Leclerc, Y. G.: Taking Advantage of Image Based and Geometry Based Constraints to Recover 3D Surfaces. *Computer Vision and Image Understanding*, Vol.64 (1996) 111-127
5. Hannah, M.: A System for Digital Stereo Image Matching. *Photogrammetric Engineering and Remote Sensing*, Vol. 55 (1989) 1765-1770
6. Panton, D. J.: A Flexible Approach to Digital Stereo Mapping. *Photogrammetric Engineering and Remote Sensing*, Vol. 44 (1978) 1499-1512
7. Cochran, S. D., Medioni, G.: 3-D Surface Description from Binocular Stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 14 (1992) 981-994
8. Mori, K., Kidode, K., Asada, H.: An Iterative Prediction and Correction Method for Automatic Stereo Comparison. *Computer Graphics and Image Processing*, Vol. 2 (1973) 393-401
9. Mostafavi, H.: Image Correlation with Geometric Distortion Part II: Effects on Local Accuracy. *IEEE Trans. Aerospace and Electronic*, Vol. 14 (1978) 494-500
10. Schultz, H., Woo, D., Riseman, E., Stolle, F.: Error Detection and DEM Fusion Using Self-consistency. *IEEE Int. Conf. on Computer Vision*, Vol. 2 (1999) 1174-1181
11. Scharstein, D, Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, Vol. 47 (2002) 7-42
12. Schultz, H.: Terrain Reconstruction from Widely Separated Images. *Proceeding of SPIE*, Vol. 2486 (1995) 113-122

Consideration of Illuminant Independence in MPEG-7 Color Descriptors

Sang-Kyun Kim¹, Yanglim Choi², Wonhee Choe³, Du-Sik Park³,
Ji-Yeun Kim¹, and Yang-Seock Seo¹

¹ Computing Lab, Digital Research Center, Samsung A.I.T.,
San 14-1, Nongseo-ri Giheung-up Yongin-si S.Korea
{skkim77, jiyun.kim, ysseo}@samsung.com

² Connectivity Lab, Digital Media center, SEC,
Metan 3, Paldal-gu Suwon-si S.Korea
yanglimc@samsung.com

³ Imaging Solution Program Team, Digital Research Center, Samsung A.I.T.,
San 14-1, Nongseo-ri Giheung-up Yongin-si S.Korea
{wonhee.choe, dusikpark}@samsung.com

Abstract. In this paper, we demonstrate the retrieval performance of the color descriptors in ISO/IEC International standard (i.e., MPEG-7 Visual) against the images taken under various lighting conditions. A simple and effective method for compensating illumination variation is proposed. The method proposed is proven to be useful by testing images from laboratory and outside taken under many different lighting conditions using cameras with different white balance settings.

1 Introduction

Every day we are facing humongous amount of digital media contents taken from digital camera, camcorder, broadcast, and Internet. Especially, the wide spread of digital cameras as well as the cell-phones embedded with a high resolution digital camera is accelerating the increase of personal digital image archives. People start to be desperate to find an easy, efficient and effective way of managing, indexing, and searching their pictures. Recent movements in multimedia society such as MPEG, JPEG, and Microsoft's new OS, Longhorn, reflect such needs and trends. Under the circumstances, there have been many on-going researches on content based image retrieval (CBIR) [6], [7]. Among them, the MPEG-7 standard is the one of the vigorous activities to accomplish the mission. The MPEG-7 standard defines a set of descriptors for visual media contents and is released its version 1 specification on May 2002 to public [8]. Currently, the MPEG-7 is working for its first amendment [9]. This paper is about the illumination invariant color descriptor, which is accepted as a new international standard with a unique functionality in the MPEG-7 visual group.

The MPEG-7 Visual color related descriptors standardized in [8] are for the similarity search and retrieval functionality. But the effect of illumination change in the visual content is not considered. Illuminant change, including the

presence of shadows, can cause a great deal of change in the color distribution of the image and often makes two images with similar scene lie far apart in the similarity search space. Related to this problem, active study has been going on for the past decades in the research area called the "color constancy" (i.e., finding an illumination independent representation of color) [1], [2], [3], [4], [5]. Application of results from these studies has been mainly tested on synthetic and laboratory setting images, while real-image-applicable-algorithms are very few [2], [5]. Various modeling and estimation techniques have been developed and they can be classified into two categories. One is to find a specific color representation that eliminates the illumination effect of the image color. The other is to detect the illumination components and use them in the matching process to obtain illumination independent matching result.

The proposed method is to normalize the illuminant (i.e., chromaticity in CIE 1931 diagram) (x, y) values of the image. There are various methods available to estimate this value. We have used a variant of the Grey-World algorithm, which is more basic and simpler compared to other sophisticated methods, such as the Gamut mapping, Color by correlation, or the Color by neural network. But those advanced methods need a priori information specific to a given data set, which is not a very desirable feature in terms of a general image searching task. Using the proposed descriptor we convert the image to have a canonical illuminant chromaticity at 6500K in the daylight locus. Then the MPEG-7 color descriptor of the converted image is extracted and matched as specified in the standard [14].

In section 2, a method how to estimate the scene illuminant and to adjust the calculated scene illuminant to a predefined canonical illuminant, is briefly explained. The experiment conditions, camera settings, illumination settings, and experimental results are demonstrated in section 3 followed by a conclusion in section 4.

2 Scene Illuminant Estimation and Compensation Algorithm

When illumination independence is required, the query and the database images are converted to new images using the Bradford transform matrix (3 x 3) calculated from the illuminant (x, y) values of the image and the (x, y) values of the point in the x-y plane with color temperature 6500K on the daylight locus. The color descriptors of the new images are then extracted using the standard and compared as is done in the ordinary matching procedure. This algorithm can be applied to any color descriptors in MPEG-7 visual since the extraction of the descriptor from the image is not changed. It is assumed that the illuminant (x, y) value is already available to calculate the Bradford transformation matrix. The overall procedure is depicted as a flow chart in Fig. 1. Many methods for estimating the scene illuminant color from a color image have been presented in the literatures. In this paper, we estimate the illuminant chromaticity of an image using a method based on the perceived illumination [11], [12], [13] and

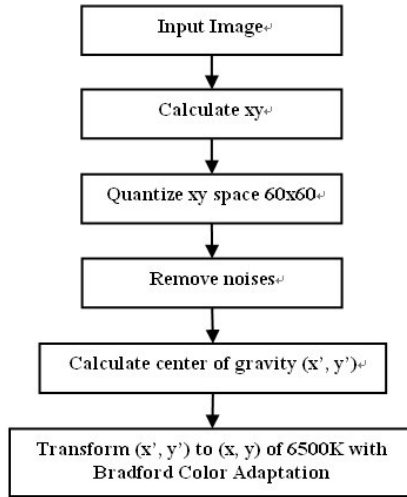


Fig. 1. Scene illuminant estimation and adjustment process flow

calculate the color temperature from the estimated chromaticity. Its detailed procedure is explained as follows:

1. The pre-processing step.
 - a) Linearizing input image: $RGB \rightarrow R_1G_1B_1$
 - b) Converting $R_1G_1B_1$ into XYZ .
 - c) Removing pixels that have the pixel value smaller than the low luminance threshold (T_{ll}).
 - d) Averaging XYZ value for all remained pixels: $X_aY_aZ_a$
 - e) Calculating the self-luminous threshold: $X_TY_TZ_T$
 - f) If $X_TY_TZ_T$ have the same values with the previous values, go to the illuminant (x, y) estimation step, else remove pixels that have the pixel value bigger than the self-luminous threshold and repeat step (d) to (f).
2. The illuminant (x, y) estimation step.
 - a) Project the remaining XYZ values to the x - y space.
 - b) The bin counting step.
 - Divide the x - y plane in 60×60 grids. We call the each cell as the 'bin'.
 - Index each bin 0 or 1 according to the presence of the (x, y) values in it, with a proper noise removal scheme. Specifically, there is a threshold value $alpha$ depending on the image size and the total bin number that, for each bin, if the number of (x, y) values belonging to that bin is less than $alpha$, then the bin index is 0, otherwise the bin index is 1.
 - c) Average the (x, y) values of each bin with index 1 to estimate the center of illuminant (x', y') . This is based on the Grey-World algorithm.

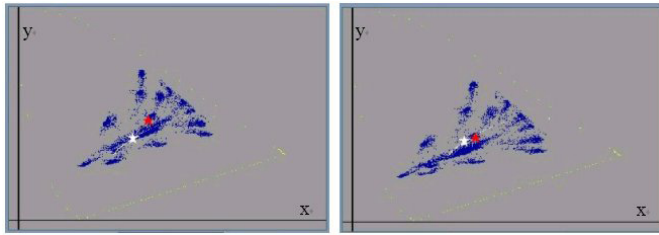


Fig. 2. Scene illuminant adjustment before (left) and after (right) using Bradford CAT

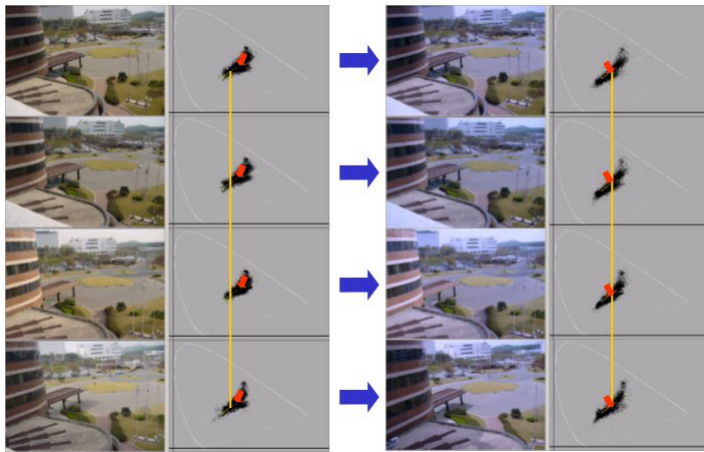


Fig. 3. Scenes before and after applying the illumination compensation method proposed. The yellow lines indicate the location of 6500K on the day light locus, while red arrows indicate the center of gravity of the scene illuminant before and after applying the algorithm proposed

3. Image conversion step: This consists of the following two steps.

- a) The image XYZ pixel value conversion by the Bradford transform [10] calculated between the illuminant (x', y') and the illuminant (x, y) of the 6500K on the daylight locus lying in the x - y plane. In Fig. 2, the red triangles indicate the location of the center of gravity of the x - y color gamut before and after applying the Bradford color adaptation transform in the $(x$ - y) chromaticity diagram. The white stars indicate the location of 6500K.
- b) Finally, convert XYZ values to RGB values.

After converting the image to the canonical illuminant 6500K, one can extract color descriptors in MPEG-7 from the converted image. Fig. 3 demonstrates the scenes before and after applying the algorithm described.

3 Experimental Results

We corrected pictures into two big sets: inside and outside. The number of images and ground truth sets, which were used, are as follows:

1. Outside natural images: 1538 images with 142 ground truth sets.
2. Images of collection of objects, taken inside: 330 images with 28 ground truth sets.

We created 1538 pictures from outside taken by Sony DSC-F505V and Samsung Digimax 350SE with the auto white balance. The size of image from Sony camera is 640x480 while the size of image from Samsung camera is 512x384. Each ground truth set contains similar contents taken by the digital camera with varying illuminant conditions of a day (i.e., noon, late afternoon, sunset, cloudy, and rainy), as well as in different zoom and rotation. We created 330 images from laboratory taken by Sony DSC-F505V with auto white balance and inside (incandescent) white balance. We used 5 different illuminants: Horizon (2300K), Illuminant A (Incandescent A: 2856K), D65 (6500K), Cool white (Fluorescent: 4150K), and Cool white (4150K) + UV (Ultrarume 30: narrow band 3000K). We composed a scene containing enough colored objects and different objects arrangement so that the distinction between scenes would be noticeable to viewers. Fig. 4 and Fig. 5 demonstrate an example of ground truth images from outside and laboratory, respectively. As noticed from Fig. 4, the outside images are not of the exactly same scene. They are taken from slightly different angles and zooms in order to account the real-life situation. Since there are enormous cases existed in the world that the same scene can be taken with different cameras with different white, we have simulated cases by using the two different cameras (Fig. 6 1st and 2nd row). Since many people uses auto or inside white balance while taking pictures inside, we have tried the two settings of the white balance control (i.e., auto and incandescent) of Sony F505V for the same inside scene (Fig. 6 3rd and 4th row).



Fig. 4. A ground truth example of outside pictures

The process flow in Fig. 7 demonstrates how to search and retrieve with a query image and DB images. The color descriptors (i.e., metadata) of images in DB should be extracted after applying the illumination compensation method proposed. The similarity measure for each MPEG-7 descriptor is well described



Fig. 5. A ground truth example of inside images (from laboratory)



Fig. 6. Outside images taken by Sony DSC-F505V with auto white balance (1st row), Outside images taken by Samsung Digimax 350SE with auto white balance (2nd row), Laboratory images taken by Sony DSC-F505V with auto white balance (3rd row), Laboratory images taken by Sony DSC-F505V with incandescent white balance (4th row)

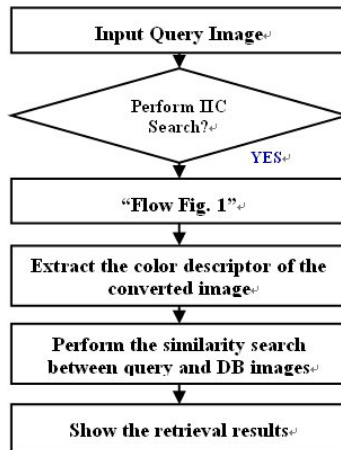


Fig. 7. A process flow explaining how to retrieve images with the illumination compensation method proposed

Table 1. ANMRR of outside image retrieval (1538 images w/ 142 GTs)

Visual Descriptor	Ordinary	Illumination Compensated	Gain	Total Size (KB)
Dominant Color	0.467572	0.322362	0.134856	24
Color Layout	0.350280	0.268037	0.087193	12
Scalable Color	0.379699	0.267750	0.111949	192
Color Structure	0.336044	0.214256	0.121788	845

Table 2. ANMRR of inside (laboratory) image retrieval (330 images w/ 28 GTs)

Visual Descriptor	Ordinary	Illumination Compensated	Gain	Total Size (KB)
Dominant Color	0.796313	0.582719	0.213594	6
Color Layout	0.680415	0.472811	0.207604	3
Scalable Color	0.720737	0.513825	0.206912	42
Color Structure	0.694470	0.458065	0.236405	83

in [14]. For example, the distance between two Color Structure descriptor histograms for image Q and I is calculated using the L_1 norm as follows:

$$D_{DCD}(Q, I) = \sum_{i=0}^{127} |(H_{Q,i} - H_{I,i})| \quad (1)$$

where $H_{Q,i}$ represents the i th bin of the color structure histogram for image Q and a 128-bin histogram has been.

Table 1 shows the Average Normalized Modified Retrieval Rank results (ANMRR) of the outside image retrieval before and after applying the illumination compensation method. The objective of ANMRR is in general to determine how much of the correct images are retrieved and how high they are ranked among the retrievals[15]. The average ANMRR from 4 MPEG-7 color descriptors is 0.2681 after applying the method, while 0.3834 before applying. The average gain is about 0.115 and this is quite substantial improvement. For the laboratory images, the average ANMRR becomes 0.5069 as in Table 2 after applying the algorithm. Even though the average gain is around 0.213, the ANMRR result, 0.5069, may not represent the acceptable retrieval performance. Any color constancy algorithm can hardly survive the extreme light changes from laboratory as shown in Fig. 6 anyhow. The Color Structure descriptor(CSD) demonstrates the best retrieval performance in overall, even though the metadata size is relatively larger than other descriptors. Since the CSD provides information regarding color distribution as well as localized spatial color structure in an image, the overall scene illuminant adjustment is presumably well suited for the CSD matching between images with a little bit of illumination shift. The Color Layout descriptor shows good results in not only the performance but also its metadata size.

4 Conclusion

We showed image retrieval performance of MPEG-7 color descriptors using images taken under various lighting conditions. We proposed a simple illumination compensation algorithm using chromaticity diagram and Grey-World theory. The algorithm is proven to be effective in the retrieval accuracy over 1538 outside images and 330 images from laboratory. The algorithm is especially effective for the outside images. This is expected to be efficient for searching ordinary photos taken from similar places with zoom and lighting changes as in the digital photo album application.

References

1. D.H. Brainard and W.T. Freeman, "Bayesian color constancy", *J. Opt. Soc. Amer.-A*, 14(7). Buluswar, S.D. Draper, "Color Constancy in outdoor images", a IEEE International Conference on Computer Vision, January 1998
2. M.S. Drew, J. Wei, and Z.-N. Li, "Illumination-invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images", *ICCV 98*, 533-540. IEEE, 1998.
3. G.D. Finlayson, S.S. Chatterjee, and B.V. Funt, "Color angular Indexing", *I ECCV96*, 11:16-27, 1996
4. D. Forsyth, "A novel approach for color constancy", *International Journal of Computer Vision*, 5:5-36, 1990
5. Yanghai Tsin, Robert T. Collins, Visvanathan Ramesh, Takeo Kanade, "Bayesian Color Constancy for Outdoor Object Recognition", *IEEE 2001 Conference on Computer Vision and Pattern Recognition*
6. A.D. Bimbo, "Visual Information Retrieval", Morgan Kaufmann Publ., Inc., 1999
7. Rui, T.S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues", *Journal of Visual Communication and Image Representation*, 10(10):39-62, 1999
8. ISO/IEC JTC1/SC29/WG11 15938-3, "Information technology - Multimedia content description interface - Part 3 Visual
9. ISO/IEC JTC1/SC29/WG11 MPEG-7 FPDAM N5695, "Text of ISO/IEC 15938-3 FPDAM 1 Visual Extensions", Trondheim, July, 2003
10. Susstrunk, S., Holm, J. and Finlayson, G.D. "Chromatic Adaptation Performance of different RGB Sensors." *Proceedings of IS&T/SPIE Electronic Imaging 2001*, Vol. 4300, 1-12, 2001
11. J.Y. Kim, S.D. Lee, C.Y. Kim, and Yang-seok Seo, "Method for determining colour of illuminant and apparatus therefor", U.S. Patent #6249601, 2001.
12. J.Y. Kim, D.S. Park, C.Y. Kim, Y.S. Seo, and Y.H. Ha, "New algorithm for detecting illuminant chromaticity from color images", *Proceedings of SPIE Electronic Imaging 2000*, Vol. 3963, 2000.
13. Du-Sik Park, Sang-Kyun Kim, Chang-Yeong Kim, Won-Hee Choi, Seong-Deok Lee, Yang-Seock Seo, "User-preferred color temperature conversion for video on TV or PC", *Proceedings of the SPIE*, Vol. 5008, 285-293, 2002
14. ISO/IEC JTC1/SC29/WG11 15938-8 PDTR "Extraction and Use of MPEG-7 Descriptions"
15. B.S. Manjunath, "Introduction to MPEG-7 : Multimedia content description interface", John Wiley & Sons (2002)

Improvement on Colorization Accuracy by Partitioning Algorithm in CIELAB Color Space

Tomohisa Takahama, Takahiko Horiuchi, and Hiroaki Kotera

Graduate School of Science and Technology, Chiba University
Chiba 263-8522, Japan,
mail-tower@graduate.chiba-u.jp
{Horiuchi, Kotera}@faculty.chiba-u.jp

Abstract. Colorization is a computerized process that adds color to monochrome images. Since different colors may carry the same luminance in spite of differences in hue and/or saturation, the colorization is an ill-posed problem. In the previous studies, one of the authors has proposed a colorization algorithm by sawing seed colors and propagating them in RGB color space. However, there is an error propagation problem at edges. In this paper, we improve on colorization accuracy for monochrome images by setting partitions which can prevent color propagation at edges. Furthermore, we newly introduce a colorization technique in CIELAB color space, and more accurate colorization can be obtained. The performance is verified by experiments.

1 Introduction

The demand of adding colors to monochrome images such as BW movies and BW photos has been increasing. For example, in the amusement field, many old movies and video clips have been colorized by human's labor, and many monochrome images have been distributed as vivid images. In other fields such as archaeology dealing with historical monochrome data and security dealing with monochrome images by a crime prevention camera, we can imagine easily that colorization techniques are useful.

A luminance value of a monochrome image can be calculated uniquely by a linear combination of an RGB color vector. However, searching for the RGB vector from a luminance value poses conversely an ill-posed problem, because there are several corresponding colors to one luminance value. Due to these ambiguous, human interaction usually plays a large role in the colorization process. The correspondence between a color and a luminance value is determined through common sense (green for grass, blue for the ocean) or by investigation. Even in the case of pseudo-coloring [1], where the mapping of luminance values to an RGB vector is automatic, the choice of the color-map is purely subjective. Since there are a few industrial software products, those technical algorithms are generally not available. However, by operating those software products, it turns out that humans must meticulously hand-color each of the individual image subjectivity. There also exist a few patents for colorization [2],[3]. However, those approaches depend on heavy human operation.

Welsh et al. proposed a coloring method using a source color image [4]. The concept of transferring color from one image to another image was inspired by work in Ref.[5]. A fast algorithm for Welshfs method has also developed [6]. In the Welshfs method, the source image, which is the same kind of image as a monochrome image, is prepared and colorization is performed by color matching between both pictures. Horiuchi et al. proposed another coloring method by sowing seed colors and propagating them [7],[8]. In those methods, miss-colorization is occurred by error propagation at edges. In this paper, we improves the colorization accuracy by modifying the colorization algorithm in Ref.[8]. At first, we develop partitioning algorithms for prevent error propagation at edges. Then we newly introduce a colorization process in CIELAB color space. By using CIELAB space instead of conventional RGB color space, we show that PSNR is improved.

This paper organized as follows: Section 2 presents a previous colorization algorithm. In section 3 we describe the proposed partitioning algorithms. In section 4 we introduce a colorization process in CIELAB color space. Section 5 shows the experimental results obtained with the proposed method. Section 6 presents conclusions.

2 Conventional Colorization Algorithm

In this section, we describe an outline of the previous colorization algorithm in Ref. [8]. Let us consider $M \times N$ pixel image. Let $N_{i,j} = (i, j), (1 \leq i \leq M, 1 \leq j \leq N)$ be an image coordinate (i, j) . Let $\mathbf{I}_{i,j}$ be an RGB color vector $(R_{i,j}, G_{i,j}, B_{i,j})$ and $Y_{i,j}$ be a corresponding luminance value. The element of $\mathbf{I}_{i,j}$ and $Y_{i,j}$ are quantized by L_1 and L_2 bits, respectively. It is well-known that a color vector $\mathbf{I}_{i,j}$ and the luminance value $Y_{i,j}$ have the following relation:

$$Y_{i,j} = (0.299, 0.587, 0.114)\mathbf{I}_{i,j}^T. \quad (1)$$

Here, symbol T expresses transposition of a matrix. The colorization problem amounts to replacing a scalar luminance value $Y_{i,j}$ stored at each pixel of a monochrome image by a three-dimensional color vector $\mathbf{I}_{i,j}$. Thus, this is an ill-posed problem for which it makes no sense to try to find an optimum solution. The aim of the colorization is how to select the suitable color for each pixel out of many candidate colors.

In Ref.[8], the following colorization algorithm was proposed.

(STEP1) Sowing seed pixels: Seed colors are set on a monochrome image. Many seeds can be set in parallel, but the behavior of one seed pixel is explained here. Sowing a seed color on $N_{i,j}$ means to choose the optimum color $\mathbf{I}_{i,j}^*$ out of candidate colors $\{\mathbf{I}_{i,j}\}$ for the luminance value $Y_{i,j}$. That is, the relation of Eq.(1) is satisfied between $Y_{i,j}$ and $\mathbf{I}_{i,j}^* \in \{\mathbf{I}_{i,j}\}$.

Figure 1(a) shows a given seed pixel.

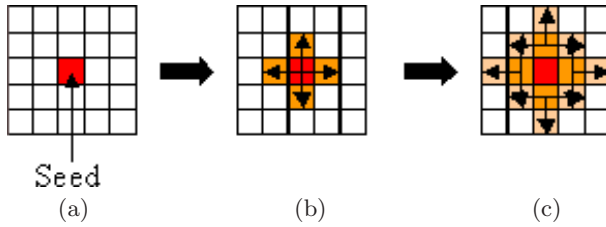


Fig. 1. The proposed propagating process: (a) A seed, (b) Initial propagation, (c) 2nd propagation.

(STEP2) Initial propagation: The seed color $I_{i,j}^*$ propagates to adjacent four-connected pixels $N_{i-1,j}, N_{i+1,j}, N_{i,j-1}, N_{i,j+1}$. For example, the optimum color in a pixel $N_{i-1,j}$ is determined by the following Eq.(2) out of the candidate color $\{I_{i-1,j}\}$.

$$J(I_{i-1,j}) = \|I_{i-1,j} - I_{i,j}^*\| \rightarrow \min. \tag{2}$$

Here, symbol $\| \cdot \|$ shows the Euclidean distance in RGB color space. Figure 1(b) shows an image after the initial propagation.

(STEP3) Iterative propagation: The propagation process proceeds to the next four connected components. In the case that colors are propagated from many adjacent pixels at the same time as shown in Fig.1(c), a suitable color is selected from candidate colors to minimize the average of color differences for each adjacent propagated pixels.

(STEP4) End conditions: (STEP3) is continued until all pixels are colored. If a propagation wave conflicts with other waves, the propagation process stops at the point.

Figure 2 shows an example of actual propagation process. Figure 2(a) shows a monochrome image. 9 color seeds were planted in the shape of lattice as shown in Fig. 2(b). From Fig. 2(c) to Fig. 2(e) show the process of colorization. Finally, we can get the colorized image as shown in Fig.2(f).

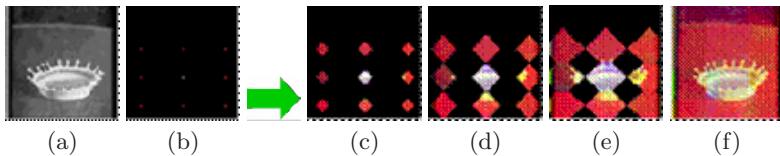


Fig. 2. Previous colorization algorithm in Ref.[8] by giving 9 seeds.

Although the method in Ref.[8] works under an assumption that natural images are smooth color change locally, the assumption is not necessarily right at edges. Therefore, there was a problem that incorrect propagation might occur

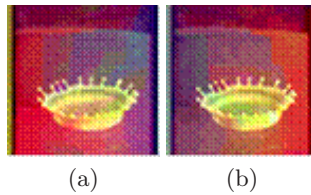


Fig. 3. Examples of miss-colored images.

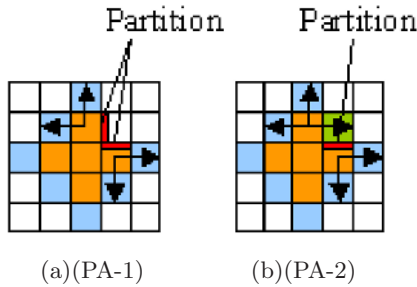


Fig. 4. Two kinds of proposed partitioning algorithms.

at edges and the error affected future propagation. Figure 3 shows examples of the colorization result by giving one seed pixel to the red background. By propagating through a crown, error propagation has arisen.

3 Proposed Partitioning Algorithms

In this section, two kinds of partitioning algorithms (PAs) are proposed to solve the error propagation problem described in the previous section.

(PA-1): When the average of minimized difference colors among all propagated pixels (STEP 3 in Sec.2) is larger than a threshold, partitions are set around the pixels and the propagation will stop as shown in Fig.4(a).

(PA-2): When a minimized difference color for a propagated pixel is larger than a threshold, a partition is set between those pixels as shown in Fig.4(b).

Figure 5 shows colorized results by setting partitions. By setting partitions, the propagation can be stopped at edges.

4 Colorization in CIELAB Space

As mentioned in Sec.2, conventional algorithm used the Euclidean distance in RGB space for the determination of the color propagation. The RGB color space was also used in the judgment of the partitioning. The RGB space is easy to

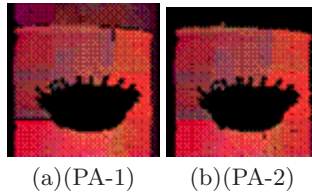


Fig. 5. Two kinds of proposed partitioning algorithms.

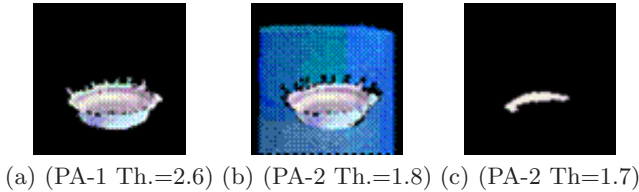


Fig. 6. Verification of partitioning algorithm in RGB space.

treat color signals, but it is not a perceptual space. Then, in this section, we propose a colorization algorithm by using a color difference ΔE_{ab}^* in CIELAB space which is a perceptual space. (STEP2) in Sec.2 is modified as follows.

(STEP2’): The optimum color for a pixel $N_{i-1,j}$ is determined by satisfying the following Eq.(3) out of the candidate color $\{vecI_{i-1,j}\}$.

$$J(I_{i-1,j}) = \|F(I_{i-1,j}) - F(I_{i,j}^*)\| \rightarrow \min. \tag{3}$$

Here, $F(\cdot)$ is the $L^*a^*b^*$ color vector of an RGB vector “.”. The symbol $\|\cdot\|$ shows the color difference ΔE_{ab}^* in CIELAB color space.

5 Experiments

In order to verify the proposed method, colorizing experiments were performed. In the experiments, the image size is set to 80×80 , and it is expressed by 4-bit of each RGB components which are the same condition in Ref.[8]. At first, we explain the effectiveness by using Milkdrop image from SIDBA database.

5.1 Verification of Partitioning Algorithms

First, in order to verify the partitioning algorithm proposed in section 3, one seed was sowed at (45,50) inside the crown and performed the colorization. The seed color was transferred from the original color image. Figure 6(a) shows the colorized result by using (PA-1). The partition was acting appropriately at edges. Figure 6(b) and 6(c) show the result by using (PA-2). In the case of Fig.6(b), the error propagation was occurred through a low partition. By changing the

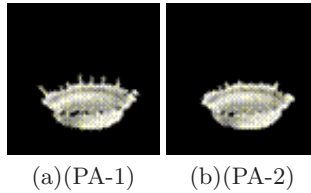


Fig. 7. Verification of partitioning algorithm in CIELAB space.

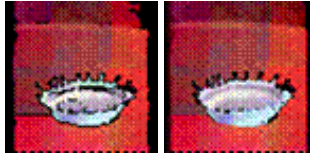


Fig. 8. Examples of colorized results with uncolorized regions.

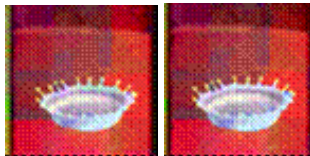


Fig. 9. Colorized results after collapsing partitions for images in Fig.8.

threshold (Th.) of partitions, seed color was propagated within the small region as shown in Fig.6(c). The result shows that the colorized result using (PA-2) is sensitive to the threshold of the partition.

Next, the experiment was performed in the CIELAB space under the same condition. The colorized result is shown in Fig.7. Appropriate and stable results were given for both partitions.

5.2 Verification of Colorization

Next, the colorization result was verified. Figure 8 shows colorized results by giving nine seeds to the rectangle of the image. There was a problem that many isolated uncolorized regions were occurred by partitions. So, we developed a collapse algorithm. In the collapse algorithm, all partitions were collapsed by setting the threshold as zero. After the colorization with partitions, the collapse algorithm is performed and all isolated regions will be colorized. Figure 9 shows colorized results after collapsing partitions for images in Fig.8.

By using the collapse algorithm, experiments on colorization were performed in both RGB and CIELAB color spaces. Nine seeds were given to the rectangle of the image, and the seed pixel was given to the same place. The PSNR in RGB

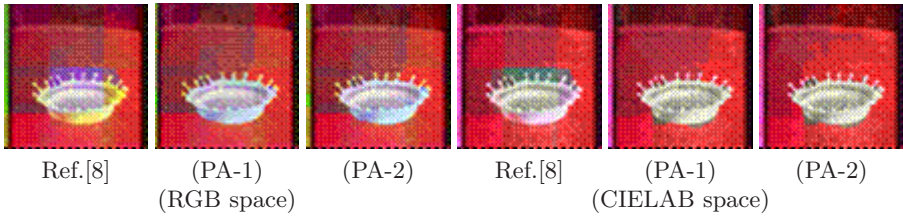


Fig. 10. Objective evaluation for colorization in RGB and CIELAB spaces.

Table 1. Objective evaluation for colorization in RGB and CIELAB spaces.

	(RGB space)			(CIELAB space)		
	Ref.[8]	(PA-1)	(PA-2)	Ref.[8]	(PA-1)	(PA-2)
PSNR[dB]	18.17	18.85	18.78	19.15	20.85	20.44
ΔE_{ab}^*	15.86	15.65	15.30	14.16	11.61	12.35

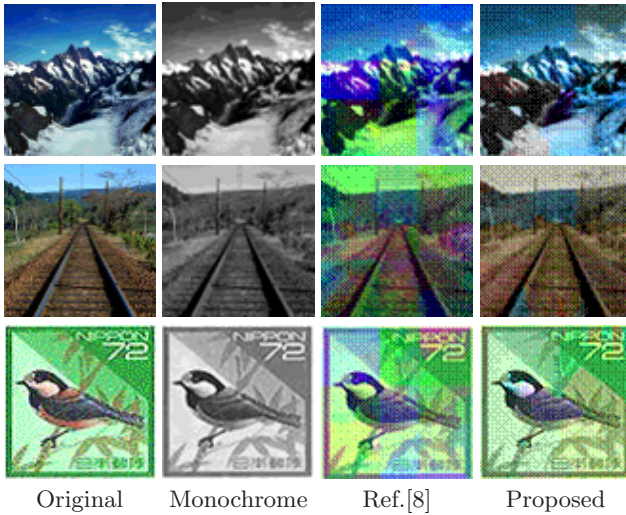


Fig. 11. Colorization results.

space and the color differences ΔE_{ab}^* in CIELAB space were used as objective evaluation. Figure 10 and Table 1 show the results.

Colorization results with the proposed partitioning algorithms obtained better rates than Ref.[8]. Moreover, the colorization in the CIELAB color space obtained better results than the colorization in the RGB color space. Other colorized results were shown in Fig.11. As the proposed algorithm in Fig.11, we performed colorization with (PA-1) in the CIELAB space. It can be confirmed

that the colorization accuracy is improved using the proposed algorithm by comparing with conventional algorithm in Ref.[8].

6 Conclusions

In this paper, partitioning algorithms are proposed to improve the colorization accuracy. Then, we showed that the accuracy can be improved more and more by using the CIELAB color space which is a perceptual color space. By using the proposed partitioning algorithms, error propagation can be prevented at edges. The problem of isolated uncolorized regions was solved by developing a collapse algorithm. By using the CIELAB color space, both subjective and objective evaluations obtained good results.

Future works are to investigate the optimum number of seeds and find out the optimum position of seeds.

References

1. Gonzales, R.C., Wintz, P.: Digital Image Processing. Addison-Wesley Publishing, (1987)
2. Wilson, M., Brian, H.: Coloring a black and white signal using motion detection. Canadian Patent, No.CA 01291260, (1991)
3. Roy, P.V.: Designing, drawing, and colorizing generated images by computer. U.S. Patent, No.5,831,633, (1998)
4. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to grayscale image. Proc. ACM SIGGRAPH2002 **20** 3 (2002) 277–280
5. Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **Sep./Oct.** (2001) 34–40
6. Blasi, G.D., Recupero, D.-R.: Fast Colorization of Gray Images. Proc. Eurographics Italian Chapter (2003)
7. Horiuchi, T.: Colorization algorithm using probabilistic relaxation. Image and Vision Computing **22** 3 (2004) 197–202
8. Horiuchi, T., Hirano, S.: Colorization algorithm for grayscale image by propagating seed pixels. Proc. IEEE International Conference on Image Processing (2003)

Gabor-Kernel Fisher Analysis for Face Recognition

Baochang Zhang

Department of Computer Science and Engineering, Harbin Institute of Technology,
Harbin, 150001, China
Bczhang@jdl.ac.cn

Abstract. Kernel based methods have been of wide concern in the field of machine learning. This paper proposes a novel Gabor-Kernel Fisher analysis method (G-EKFM) for face recognition, which applies Enhanced Kernel Fisher Model (EKFM) on Gaborfaces derived from Gabor wavelet representation of face images. We show that the EKFM outperforms the Generalized Kernel Fisher Analysis (GKFD) model. The performance of G-EKFM is evaluated on a subset of FERET database and CAS-PEAL database by comparing with various face recognition schemes, such as Eigenface, GKFA, Image-based EKFM, Gabor-based GKFA, and so on.

Keywords: Gabor, Kernel Fisher, Face Recognition.

1 Introduction

Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two classical techniques for linear feature extraction. Although LDA is conceptually simple and has been used in many applications, it has some limitations: it requires at least one of scatter matrices to be nonsingular and it can not easily capture a nonlinearly clustered structure. In recent years, the nonlinear feature extraction methods, such as Kernel Principal Component Analysis (KPCA) and Kernel Fisher Discriminant Analysis (KFD) have been of wide concern. KPCA was originally developed by Scholkopf [1], and KFD was subsequently proposed by Mika [2] and Baudat [3]. However, the KFD always faces the difficulty in its application. The reason is that KFD is trained by using mapped training samples, which make within-class covariance matrix singular. Moreover, when the input space is mapped to a feature space through a kernel function, the dimension of the feature often becomes larger than that of the sample space, and as a result, the scatter matrices become singular. In order to solve this problem, we proposed Enhanced Kernel Fisher Model (EKFM), which never faces the difficulty of calculation of the inverse of the within-class.

In this paper, we apply the proposed scheme to face recognition issue, which is one of hot points in the field of pattern recognition. A good face recognition methodology should consider representation as well as classification issues, and a good representation method should require minimum manual annotations[4,5,7].

The Gabor wavelets, whose kernels are similar to the 2D receptive field profiles of the mammalian cortical simple cells, exhibit desirable characteristics of spatial locality and orientation selectivity[6]. The Gabor wavelet representation facilitates recognition without correspondence, because it captures the local structure corresponding to spatial frequency (scale), spatial localization, and orientation selectivity [6,7], which is augmented into Gaborfaces by concatenating various scales and orientations [7,8]. Here, we will give the brief organization of this paper. In part 2, the Gaborface feature is briefly introduced, which focuses on the representation of face image. In part 3, G-EKFM is proposed by using the Enhance Kernel Fisher Model. In part 4, we will give some experiment results on CAS-PEAL and FERET Databases. In the last part, we will make some conclusions about the experiments results.

2 Gabor Wavelet Representation

Gabor wavelet model quite well the receptive field profiles of cortical simple cells. Lades et al.[9] pioneered the use of Gabor wavelet for face recognition using the Dynamic Link Architecture framework. Wiskott et al.[10] developed a Gabor wavelet based elastic bunch graph matching method to label and recognize human faces. M.J.Lyons [11] had shown through experiments that the Gabor wavelet representation is optimal for classifying facial actions.

2.1 Gabor Wavelet

Daugman pioneered the using of the 2D Gabor wavelet representation in computer vision in 1980's [6]. The Gabor wavelet representation allows description of spatial frequency structure in the image while preserving information about spatial relations [6,8,11]. A complex-valued 2D Gabor function is a plane wave restricted by a Gaussian envelope:

$$\psi_{u,v}(z) = \frac{\|k_{u,v}\|^2}{\sigma^2} * \exp(-\|k_{u,v}\|^2 * \|z\|^2 \setminus (2 * \sigma^2)) * [\exp(ik_{u,v}z) - \exp(-\delta^2)] \quad (1)$$

Here 5 frequencies and 8 orientations are used, Fig.1 shows the 40 Gabor Kernels in Eq.1 used by us.

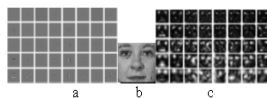


Fig. 1. a is the Real Part of 40 Gabor Kernels, b is the face image, c is Magnitude of Gaborfaces

2.2 Gaborfaces

Here, we will give a brief description about the Gaborfaces, the details of which are shown in [7,8]. In a given image, the convolution can be defined as

$$G_{u,v}(z) = I(z) * \psi_{u,v}(z) \quad (2)$$

where $*$ denotes the convolution operator, and $G_{u,v}(z)$ is the convolution result corresponding to the Gabor kernel at scale u and orientation v , named by Gaborface (shown in Fig.1). In order to utilize different spatial frequencies, spatial localities, and orientation selectivity, we concatenate all these representation results and derive an augmented feature vector. Now, $G_{u,v}^N(z)$ is the normalized vector constructed from the Gabor feature vector (Gaborface), and then x^N is defined as following:

$$x^N = (G_{0,0}^N(z), G_{0,1}^N(z), \dots, G_{4,7}^N(z)) \quad (3)$$

The dimension of Gaborfaces is very high, In order to reduce the dimensionality, at the same time reserve the identification information, the Principal Component Analysis (PCA) method is used here.

$$\Sigma_x = E((x - E(x))(x - E(x))^T) \quad (4)$$

The PCA of a random vector x factorizes its covariance matrix, then get the transform matrix P , which is an orthogonal eigenvector matrix. An important property of PCA is its optimal signal reconstruction in the sense of minimum mean-square error when only subsets of principal components are used to represent the original signal. Following this property, an application of PCA is dimensionality reduction.

$$y^p = P^m x^N \quad (5)$$

The lower dimension vector y^p captures the most expressive features of the original data.

3 Gabor-Kernel Fisher Analysis

We describe in this part the Enhance Kernel Fisher Discriminant Model and Gabor feature based kernel Fisher analysis, which are main contributions of our paper.

3.1 Kernel Fisher Analysis

The idea of Kernel FDA is to yield a nonlinear discriminant analysis in the higher space. The input data is projected into an implicit feature space by nonlinear mapping, $\Phi : x \in R^N \rightarrow f \in F$, then seek to find a nonlinear transformation matrix, which can maximize the between-class scatter and minimize the within-class scatter in F [12, 13]. First, we define the dot product in F as following.

$$k(x, y) = \Phi(x) \cdot \Phi(y) \quad (6)$$

Between-class scatter matrix S_B and Within-class scatter matrix S_w are defined in the feature space F :

$$S_w = \sum_{i=1}^C p(\omega_i) E((\Phi(x) - u)(\Phi(x) - u)^T) \tag{7}$$

$$S_B = \sum_{i=1}^C p(\omega_i)(u_i - u)(u_i - u)^T \tag{8}$$

u_i Denotes the sample mean of class i and u denotes mean of all the samples in F , $p(\omega_i)$ is the prior probability. To perform LDA in F , it is equal to maximize the following equation

$$J(w) = \frac{wS_Bw^T}{wS_ww^T} \tag{9}$$

Because any solution w must lie in the span of all the samples in F , there exists:

$$w = \sum_{i=1}^n \alpha_i \Phi(x_i) \tag{10}$$

Then maximizing Eq.9 is converted to maximize Eq.11

$$J(w) = \frac{wK_Bw^T}{wK_ww^T} \tag{11}$$

Details of K_B, K_w the can be seen in [3,12], Similar to LDA, this problem can be solved by finding the leading eigenvectors of $(K_w)^{-1}K_B$ showed in Liu[12] and Baudat(GDA)[3], which is the Generalized Kernel Fisher Discriminant(GKFD). In our paper, using the technique of pseudo inverse of the within-class covariance matrix, and the projection of a point x onto w in F given by:

$$w.\Phi(x) = \sum_{i=1}^n \alpha_i k(x_i, x) \tag{12}$$

3.2 Enhanced Kernel Fisher Model

We will propose the Enhanced Kernel Fisher (EKFM) model, which is more effective than the GKFD. Within-class covariance may be ill-conditioned, and so a regularized solution is obtained by substituting $K_w = K_w + \mu I, \mu$ is a regularization constant. The EKFM improves the generalization capability by decomposing its procedure into a simultaneous diagonalization of the two within- and between-class scatter matrices. The simultaneous diagonalization is stepwisely equivalent to two operations, and we can refer to [14, 15]. Especially in [15], Liu showed that the Enhanced Fisher Models achieved better performance than LDA. Our ideas partly originate from his methods. First we will whiten the within-class scatter matrix as following:

$$k_w \Xi = \Xi \ell, \Xi \Xi^T = I \tag{13}$$

where Ξ, ℓ are the eigenvector and the diagonal eigenvalue matrices of k_w . Ξ^*, ℓ^* are calculated by reserving $\min(l, c-1)$ eigenvectors and corresponding diagonal eigenvalue matrix, l is the length of input vector, and c is the number of classes. Then we proceed to compute the new between-class scatter matrix by using following method:

$$\ell^{*-1/2} \Xi^{*T} K_B \Xi^* \ell^{*-1/2} = \Xi_B \tag{14}$$

Diagonalizing now the new Between-class scatter matrix.

$$(\Xi_B)\Theta = \Theta\gamma, \Theta\Theta^T = I \tag{15}$$

where Θ, γ are the eigenvector and the diagonal eigenvalue matrices of Ξ_B . The overall transformation matrix is now defined as follows.

$$\alpha = \Xi^* \ell^{*-1/2} \theta \tag{16}$$

Here, we can get w as the transform matrix by using the Eq.16, and the kernel feature is calculated by using Eq.17.

$$v = w \cdot \Phi(x) = \sum_{i=1}^n \alpha_i k(x_i, x) \tag{17}$$

3.3 Similarity Measure for G-EKFM

When an image is presented to the proposed method, the augmented Gabor feature vector of the image is first calculated by using Eq.3 as detailed in section 2.2 and the lower dimensional feature, y^p , is derived by using Eq.5. The dimensionality of the discriminant feature space is determined by the Enhanced Kernel Fisher method, as defined by Eq.18. The new feature vector, v , of the image is defined as following:

$$v = w \cdot \Phi(y) = \sum_{i=1}^n \alpha_i k(y_i, y) \tag{18}$$

Given that v_1, v_2 are the extracted feature vectors corresponding to two face images x_1, x_2 . The similarity rule is based on the cross correlation between corresponding vectors.

$$d(x_1, x_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|} \tag{19}$$

Experiments are performed on two databases, CAS-PEAL and FERET databases. In our paper, the kernel function is polynomial used in the proposed EKFM method, $k(x, y) = (\frac{x \cdot y}{c} + 1)^r$, r is a constant, which is related to the length of the input vector.

4 Experiment

In our experiments, the face image is cropped to size of 32X32 and overlapped with a mask to eliminate the background and hair. For all images concerned in the experiments, no preprocessing is exploited.

Table 1. Experiments on selection of parameter for polynomial kernel function

	I-EKFM		G-EKFM	
	r = 1	r = 2	r = 1	r = 2
Accessory	53.4	62.8	72.2	72.9
Background	94.1	93.4	93.6	93.4
Distance	89.1	93.8	98.1	99.6
Expression	60.8	77	90.8	91.9
Lighting	15.1	14.1	14.4	17.6
Aging	69.7	77.3	90.9	92.4

Table 2. Experiments Results on the CAS-PEAL database(r = 2)

	Eigenface	I-GKFD	G-GKFD	I-EKFM	G-EKFM
Accessory	37.1	54	70.2	62.8	72.9
Background	80.4	96.1	90.2	93.4	93.4
Distance	74.1	93	99.6	93.8	99.6
Expression	53.6	72	89.8	77	91.9
Lighting	8	9	11	14.1	17.6
Aging	50	63.5	87.8	77.3	92.4

4.1 CAS-PEAL Database

The CAS-PEAL face database was constructed under the sponsors of National Hi-Tech Program and ISVISION. Currently, the CAS-PEAL face database contains 99,594 images of 1040 individuals with varying Pose, Expression, Accessory, and Lighting (PEAL). In this experiment, only one face image for each person was used as Gallery database. Details about this database are shown in <http://jdl.ac.cn>.

In this part, we will give some experiments results about the selection of parameter of the polynomial kernel function. Note that the first order polynomial is equivalent to the LDA method. Therefore, the kernel method is better than LDA approach. I-GKFD is the image based GKFD method, G-GKFD is the Gabor based GKFD method, and I-EKFM is the image based EKFM method. the experiment results are shown in Table.1 and Table.2.

4.2 FERET Database

The proposed algorithm is also tested on a subset of the FERET face image database. This subset includes 1400 images of 200 individuals, and each individual has 7 images. It is composed of images named with two-character strings, "ba", "bj", "bk", "be", "bd", "bf" and "bg". This subset involves variations in facial expression, illumination, and pose. The accurate rate in the Table.3 and Table.4 is the average one, where we divide 200 people into two subsets, one of

Table 3. Experiments on selection of parameter for polynomial kernel function

Methods	r = 1	r = 2
I-EKFM	76.2	85.1
G-EKFM	87	90.5

Table 4. Experiment Results on FERET database($r = 2$)

Eigenface	I-GKFD	G-GKFD	I-EKFM	G-EKFM
37.1	81	83.3	85.1	90.5

which is used to train the EKFM model and the rest is used to test the proposed approach. Both subsets have 100 people, with 7 pictures for each one. In our experiments, only “ba” part was used as gallery database, others are probe databases. Thus, we can do two groups of experiments by using this kind of partition method of the FERET database. So the results are the average accurate rate of two groups of experiments.

5 Conclusions and Future Work

We have introduced in this paper Gabor feature based Kernel Fisher Analysis method for face recognition. The Enhance Kernel Fisher model achieves better performance than the GKFD method, and the experiments are performed on the FERET and CAS-PEAL databases. The Gabor transformed face images yield features that display scale, locality, and orientation selectivity. The effectiveness of the method is shown in terms of both absolute performance indices and comparative performance against various approaches such as Eigenface, EKFM, Generalized Kernel Fisher Analysis method, and Gabor based GKFD and son on. The excellent performance shown by the method is the direct result of coupling an augmented Gabor feature vector with the EKFM method.

Our next goal is to further search for an optimal and sparse code resulting from the Gabor wavelet representation of face images, for example, AdaBoost method, before forming the augmented Gabor feature vector and applying the G-EKFM method for classification. Another possibility is that we can try to use the support vector machines to train the classifier and aim to get more generalized classifier.

Acknowledgements. This research is partially supported by National Hi-Tech Program of China (No.2001AA114190 and No. 2002AA118010), National Nature Science Foundation of China (No. 60332010), and ISVISION Technologies Co. Ltd. Specially, many thanks for the advice from Dr. Shiguang Shan.

References

1. B.Scholkopf, A.Smola, K.R.Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* 10(5) (1998) 1299-1319.
2. S.Mika, G.Ratsch, J.Weston, B.Scholkopf, K.R.Muller, Fisher discriminant analysis with kernels *IEEE International Workshop on Neural Networks for Signal Processing*, Bol.IX, Madison, USA, August, 1999, pp.41-48.
3. G.Baudat, F.Anouar, Generalized discriminant analysis using a kernel approach, *Neural Computation* 12(10) (2000) 2385-2404.
4. R.Chellappa, C.L.Wilson, and S.Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, 1995.
5. J.G.Daugman, "Face and gesture recognition: Overview," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 675-676, 1997.
6. J.G.Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Research*, vol. 20, pp. 847-856, 1980.
7. Chengjun Liu and Harry Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition" *IEEE Trans. Image Processing* vol.11 no.4 2002.
8. Baochang Zhang, Wen Gao, Shiguang Shan, Yang Peng. Discriminant Gaborfaces and Support Vector Machines Classifier for Face Recognition. *Asian Conference on Computer Vision. ACCV2004*, Jeju Island, Korea, Jan.27-30, 2004, pp37-42.
9. M.Lades, J.C.Vorbruggen, J.Buhmann, J.Lange, C.von der Malsburg, Wurtz R.P., and W.Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Computers*, vol. 42, 1993.
10. L.Wiskott, J.M.Fellous, N.Kruger, and C.von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, 1997.
11. M.J.Lyons, J.Budynek, A.Plante, and S.Akamatsu, "Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis," in *Proc. the fourth IEEE international conference on automatic face and gesture recognition*, 2000.
12. Qingshan Liu, Rui Huang, "Face Recognition Using Kernel Based Fisher Discriminant Analysis" *FGR2002*.
13. K.Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, second edition, 1991.
14. Chenjun Liu, Harry Wechsler, Enhanced Fisher Linear Discriminant Models for face Recognition, *14th International Conference on Pattern Recognition, ICPR'98*.

Film Line Scratch Detection Using Neural Network

Sin Kuk Kang¹, Eun Yi Kim², Kee Chul Jung³, and Hang Joon Kim¹

¹ Department of Computer Engineering, Kyungpook National Univ., Korea
{skkang,hjkim}@ailab.knu.ac.kr

² Department of Internet and Multimedia Engineering, Konkuk Univ., Korea
eykim@konkuk.ac.kr

³ School of Media, College of Information Science, Soongsil Univ., Korea
kcjung@ssu.ac.kr

Abstract. Line scratches are one of the most common degradations in old films. To support a demand of high quality of multimedia service, these should be detected and restored automatically. However, although many detection and restoration algorithms have been researched, little have done in automatic scratch detection. This paper presents a texture-based object detection method for scratch detection. We use a multi-layer perceptron (MLP) to automatically make a texture classifier that discriminates between scratch regions and non-scratch ones in various environments. To assess the validity of the proposed method, it has been tested with all kinds of scratches, that is, principal/secondary scratches, alone/not-alone ones, and moving/static ones, and then experimental results show that the proposed approach leads to not only robust but also efficient scratch detection.

Keywords: Neural Network Film Restoration, Scratch Detection, Texture Discrimination.

1 Introduction

In the recent years, this film restoration has gained increasing attention by many researchers. With the emergence of digital television broadcasting and the growth in video sales, there is a growing need to supply more retouchable material of an acceptable quality in a multimedia context. So a lot of work has been done for film restoration [1,2,8]. In particular a number of methods have been presented for processing the line scratches [3,4,5,6,7].

In general, an old film is degraded by dust, scratch, flick, and so on. Among these, the scratches, which are the most frequent degradation, are usually generated by mechanical rubbings during a film copy and appear in the direction of the film strip on successive frames over the film. They are easily visible as vertical lines of bright or dark intensity, oriented vertically over much of the images. An example of an image affected by scratches is shown in Fig. 1. Figs. 1(a) and (b) include a positive scratch showing as a black line and a negative

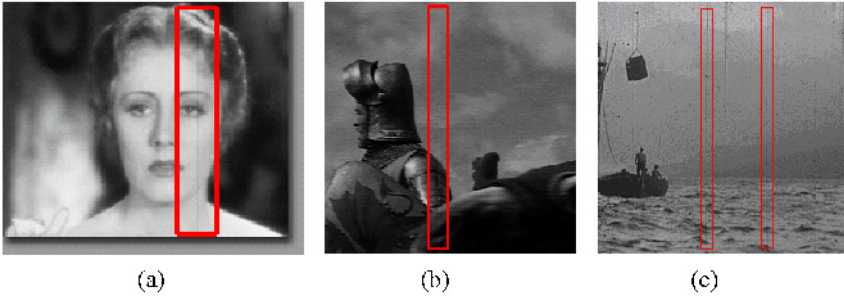


Fig. 1. Example of scratches: (a) “Stingaree de Vrijbuitter” (William Wellmann, 1934), (b) “Knight sequence”, (c) “Sitdownsequence” (The scratches are marked with gray rectangles.)

scratch showing as a white line, respectively. And a severely many scratches are shown in Fig. 1(c).

Until now, although a wide variety of literature has been written about scratch restoration, little has been done in regards to scratch detection. Moreover, the previous scratch detection techniques do not permit the detection of all kinds of scratches: most of them are restricted to principal, alone and static scratch.¹ The system should be developed that can detect all kinds of scratches from a given frame of old films.

In this paper, a neural network-based method is proposed for line scratch detection in old film archives. First, scratch pixels are identified by analyzing the textural properties of video frames. An input frame is segmented into scratch and non-scratch classes using a two-layer feed-forward neural network classifier which receives the gray values of a given pixel and its neighbors as input. The activation values of the output nodes are used to determine the class of a given central pixel. As a result of this classification, a classified image is obtained as a binary image in which the scratch pixels are black. Each pixel in the video frame belongs to one of *scratch class* and *non-scratch class*. During the post-processing step, the classified image is filtered so that sparsely distributed scratch pixels and non-vertical pixels are eliminated from the scratch class.

The paper is organized as follows: In Section 2 texture-based classification for line scratches detection is proposed. In Section 3 we present the post-processing of the neural network result. The experimental results are shown in the Section 4. Finally Section 5 draws the conclusions and future work.

2 Texture-Based Classification

A neural network is used to classify the pixels of image. After classification, a classified binary image is obtained in which scratch pixels are black. Generally, scratches have the following properties:

¹ All these terminologies are described in Section 4.

- (1) It has lower or higher brightness than neighboring pixels in its vicinity, which is used as an important cue for scratch detection.
- (2) It usually appears as a vertically long thin line.
- (3) It has temporal continuity, that is, it appears on the successive movie frames.

These properties help reduce the complexity of the problem, and facilitate the discrimination between scratch and non-scratch regions. In this work, we use the properties (1) and (2) for scratch detection. The following section describes the architecture of a neural network-based classifier and its method of classification.

2.1 Neural Network Architecture

A neural network is used as a classifier to identify scratch pixels. Scratch pixels are classed as scratch class and all other pixels are grouped as non-scratch. The input layer of the network has L_0 nodes, the hidden layer has L_1 nodes, and the output layer has 2 nodes. The adjacent layers are fully connected. The hidden layer operates as a feature extraction module. The output layer is used to determine the class of a pixel: scratch or non-scratch. A diagram of the neural network-based classifier is shown in Fig. 2.

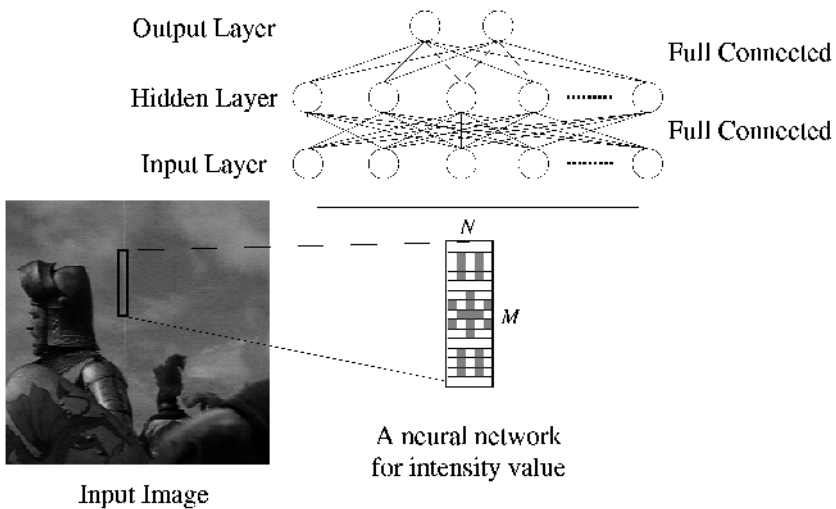


Fig. 2. A two-layer feedforward neural network.

The input layer receives the intensity values of pixels, at predefined positions inside an $N \times M$ window over an input image. The output value of a hidden node is obtained from the dot product of the vector of input values and the vector of the weights connected to the hidden node. It is then presented with the output nodes. The weights are adjusted by training with a back-propagation algorithm in order to minimize the sum of squared error during the training session.

2.2 Classification

Neural networks are used as filters which produce a local window-based classification of image pixels in ‘scratch’ and ‘non-scratch’ by analyzing texture properties of the sub-region of input images.

During training session, a set of patterns is used to train the weights of the network. Each pattern consists of the intensity values of a pixel and its neighbors, along with the actual class of the pixel. A class representation is given as a vector of two floating-point numbers, ranging from 0 to 1, rather than an identifier. The first value will be larger than the second value for scratch class pixels, and less than the second value for non-scratch pixels. A pattern is fed to the input layer, and the activation values of the output nodes are then used to determine the class of the pixel. The class of the winner among the output nodes becomes the class of the central pixel inside the $N \times M$ window from which the pattern is obtained. This result is then compared with the actual class presented. A result will be regarded as an error if the class reported by the classifier differs from the actual class. The weights are finally trained by minimizing the error rate.

For experiments, a set of patterns from an input frame is sent through a network. After the pattern feed-forward passes the network, the two output values are compared with each other, and the class of each pixel is determined. The class of falsely classified pixels can be altered during the post-processing step.

As a result of classification, a classified image is obtained. The classified image is a binary image in which the pixels classified as scratch are black whereas those classified as non-scratch are white. Fig. 3 shows an example of classification.

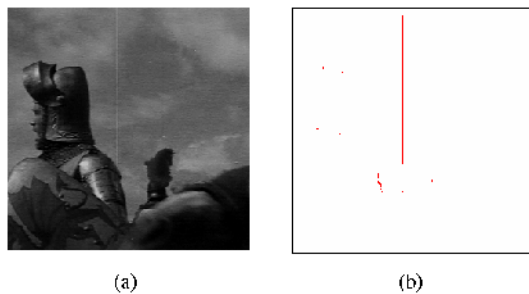


Fig. 3. Example of classification: (a) an original frame, (b) classified image.

For convenience of reporting errors, the actual class of all the pixels in each image in our database was manually labeled by marking the coordinates of all the text rectangles. Classification errors were then automatically computed by comparing the network’s output with the actual labeled class corresponding to each pixel. The value of classification error is the proportion of falsely discriminated pixels to all pixels.

3 Post-processing

The grey level of each pixel represents the output value from the neural network. Neural networks are used to produce a local window-based classification of each image pixels. So post-processing of these outputs is needed. The scratch regions are identified by filtering a classified image, using the length and direction of the scratch pixels, which are extracted by the neural network.

Some pixels may be determined as belonging to a different class from their actual class. Filtering is accomplished as follows. A pixel's class is determined as non-scratch if the number of pixels, which are classified as scratch, in the neighborhood of a 3×3 size is less than a threshold. Otherwise, the pixel's class is determined as scratch.

In addition, we have used the following three heuristics to remove text segments that hardly include characters:

- (1) The height of a scratch should be larger than 1/10 of the height of the image.
- (2) The direction of a scratch should be vertical.

The segments violating (1) and (2) are not scratch segments any more. Fig. 4 shows the post-processing results of Fig. 3(b).

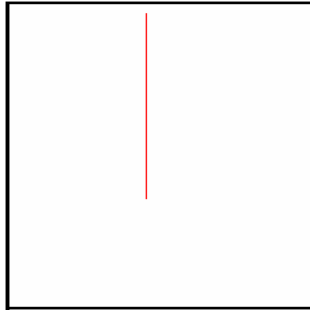


Fig. 4. Example of post-processing.

4 Experimental Results

The proposed scratch detection method was applied to some old film archives. Here, we show the results for the sequence, “Knight,” “Star” and “Sit-down.” These sequences include a variety of scratches. The scratches are divided into several types according to the length and position that appears on the consecutive frame. All types of scratches are illustrated in Table 1. For experiments, 100 frames with scratches were manually selected from these sequences. Of these frames, 50 were used in the training process, and the others were used in the

Table 1. Event detection conditions

Kinds of scratches	Description
<i>Static</i> scratches	present at the same position on consecutive frames
<i>Moving</i> scratches	can change positions during the sequence
<i>Principal</i> scratches	occurs on more than 95% of the image height
<i>Secondary</i> scratches	the others
<i>Alone</i> scratch	occurs without other scratches nearby it
<i>Not-alone</i> scratch	occurs with other scratches nearby it

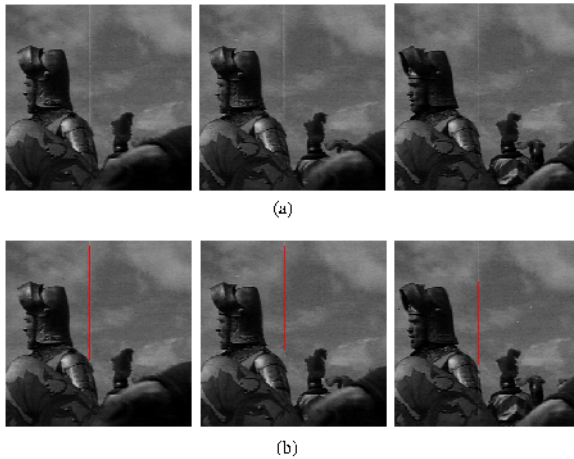


Fig. 5. Scratches detected from the sequence “Knight”: (a) input frames at time 38, 39 and 40, (b) detected scratch regions.

testing process. For each frame, a set of rectangles representing scratch regions was tagged. This was used to compare the output of the classifier with the actual class of the pixels. The neural network used in the experiments had 75 input nodes, 18 hidden nodes, and 2 output nodes. The pattern fed to the classifier consisted of 75 pixel intensity values, i.e. a central pixel and 74 neighbors, in a 15 15 window. Figs. 5-7 show the examples of the scratch detection. In Fig. 5, some frames have negative scratches. The scratches are static ones, and most of them are secondary. Fig. 5(a) shows the input frame at time 38, 39, and 40. Then the scratches detected from each frame are shown in Fig. 5(b). As can be seen in Fig. 5, the proposed method can accurately and automatically detect the static and secondary scratches.

Fig. 6 shows the scratch detection results for the sequence “Star”, which includes a number of secondary scratches. Fig. 6(a) shows an input frame with not-alone and secondary scratches, and then the classified image by a neural network

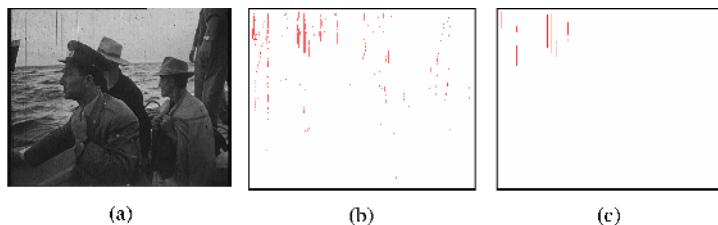


Fig. 6. Scratches detected from the sequence “Star”: (a) an input frame, (b) a classified image by a neural network, (c) a post-processing result.

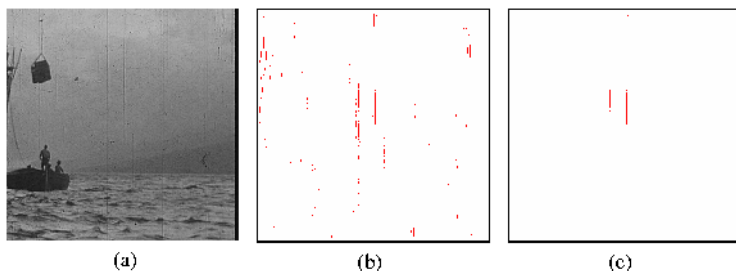


Fig. 7. Scratches detected from the sequence “Sit-down”: (a) an input frame, (b) a classified image by a neural network, (c) a post-processing result.

is shown in Fig. 6(b). Finally, the scratches regions filtered by a post-processing are shown in Fig. 6(c). As mentioned before, most of methods in the literature are limited to detect only alone and principal scratches. Unlike them, the proposed method can correctly detect these scratches, as can be seen in Fig. 6.

Fig. 7 shows the scratch detection results for the sequence “Sit-down.” The scene includes some secondary scratches and some principal ones in a frame of the sequence. Moreover, the scenes are severely degraded by other degradations, not only by scratches. Fig. 7(a) shows an input frame with not-alone and secondary scratches, and then the classified image by a neural network is shown in Fig. 7(b). Finally, the scratches regions filtered by a post-processing are shown in Fig. 7(c). In the experiments, some scratched are missed. This is due to the low contrast between real scratches and the neighboring background and low resolution of the input frame.

These experimentation shows that the proposed method can detect a variety of scratches, from alone and principal scratches to not-alone and secondary scratches. Nonetheless the proposed method may lose some not-alone scratches, it performs better than the existing techniques, as the previous methods work on only alone scratches. Moreover the proposed method is a fully automated scratch detection method with no prior knowledge.

5 Conclusion

An automatic scratch detection method for old film archives using a neural network was proposed and implemented. The main advantage of the proposed method includes the detection of not-alone and secondary scratches, as well as alone and principal scratches. In addition, the proposed method can detect line scratches in each frame without considering any knowledge on the other frames of the sequence. Although the experiment shows the effectiveness of the proposed method, it has some problems. For example, it may miss some not-alone scratches due to the low resolution of the movie films and the low contrast between scratches and the neighboring regions. Future work is focused on solving this problem.

Acknowledgements. This work is supported by Korea Research Foundation Grant (KRF-2003-042-D00166)

References

1. Schallauer, P., Pinz, A. and Hass, W.: Automatic restoration algorithms for 35mm film. *VIDERE: J. Comput. Vis. Res.*, vol. 1, No. 3, (1999)
2. Joyeux, L., Boukir, S. and Besserer, B.: Film line scratch removal using kalman filtering and Bayesian restoration. *WACV2000*, Palm Springs, CA, Dec. (2000)
3. Bruni, V. and Vitulano, D.: A generalized model for scratch detection. *IEEE Transactions on Image Processing*, Vol. 13, No. 1 (2004) 44-49
4. Joyeus, L. et al.: Film line scratch removal using Kalman filtering and Bayesian restoration. *IEEE Workshop on the Application of Computer Vision*. (2000)
5. Tegolo, D. and Isgro, F.: Scratch detection and removal from static images using simple statistics and genetic algorithms. *IEEE ICIP'2001*. (2001) 265-268
6. Maddalena, L.: Efficient methods for scratch removal in image sequences. *IEEE ICIP'2001*, (2001) 547-552
7. Kokaram, A. C.: Detection and removal of line scratches in degraded motion picture sequences. *Signal Processing*, Vol. 1. (1996) 5-8
8. Kokaram, A. C.: *Motion Picture Restoration: Digital Algorithms for Artifact Suppression in Degraded Motion Picture Film and Video*. Springer-Verlag. (1998)

A Simplified Half Pixel Motion Estimation Algorithm Based on the Spatial Correlation

HyoSun Yoon¹ and Miyoung Kim^{2,*}

¹ Department of Computer Science, Chonnam National University, 300
Youngbong-dong, Buk-gu, Kwangju 500-757, Korea
estherymoon@hotmail.com

² Department of Computer Information Technology, Namdo Provincial College, 262
Hanggyo-Ri, Damgyang-Gun, Chonnam-Provice 517-802, Korea
kimmee@namdo.ac.kr

Abstract. Motion estimation which consists of integer pixel motion estimation and half pixel motion estimation is very computationally intensive part. To reduce the computational complexity, many methods have been proposed in both integer pixel motion estimation and half pixel motion estimation. For integer pixel motion estimation, some fast methods could reduce their computational complexity significantly. There remains, however, room for improvement in the performance of current methods for half pixel motion estimation. In this paper, a simplified half pixel motion estimation algorithm based on spatial correlations is proposed to reduce the computational complexity. According to spatially correlated information, the proposed method decides whether half pixel motion estimation is performed or not for the current block. Experimental results show that the proposed method outperforms most of current methods in computational complexity by reducing the number of search points with little degradation in image quality.

1 Introduction

Motion estimation (ME) and motion compensation techniques are an important part of video encoding systems, since it could significantly affect the compression ratio and the output quality. But, ME is very computational intensive part.

Generally, ME is made of two parts, integer pixel motion estimation and half pixel motion estimation. For the first part, integer pixel motion estimation, many search algorithms such as Diamond Search (DS) [1,2], Three Step Search (TSS) [3], New Three Step Search (NTSS) [4], Four Step Search (FSS) [5], Two Step Search (2SS) [6], Two-dimensional logarithmic search algorithm [7], HEXagon-Based Search (HEXBS) [8], Motion Vector Field Adaptive Search Technique (MVFAST) [9] and Predictive MVFAST (PMVFAST) [10] have been proposed to reduce the computational complexity. Some algorithms among these algorithms can find an integer pixel Motion Vector (MV) by examining less than 10 search

* Corresponding author

points. For the second part, half pixel motion estimation, Full Half pixel Search Method (FHSM) examines eight half pixel points around the integer motion vector. This method takes nearly half of the total computations in the ME that uses fast algorithms for integer pixel motion estimation. Therefore, it becomes more important to reduce the computational complexity of half pixel motion estimation. For these reasons, Horizontal and Vertical Direction as Reference (HVDR) [11], the Parabolic Prediction-based, Fast Half Pixel Search algorithm (PPHPS) [12] and Chen's Fast Half Pixel Search algorithm (CHPS)[13] have been proposed. Since these algorithms do not have any information on the motion of the current block, they always perform half pixel motion estimation to find a half pixel motion vector.

In this paper, we propose a simplified algorithm based on spatial correlations for half pixel motion estimation. According to the information of spatially correlated motion vector, the proposed method decides whether half pixel motion estimation is performed or not for the current block.

This paper is organized as follows. Section 2 describes the previous works. The proposed method is described in Section 3. Section 4 reports the simulation results and conclusions are given in Section 5.

2 The Previous Works

In Motion Estimation and Compensation, half pixel motion estimation is used to reduce the prediction error between the original image and the predicted image. FHSM examines eight half pixel points around the integer motion vector 'C' illustrated in Fig. 1. The point with the minimum cost function value among these points is the half pixel motion vector. To reduce the computational complexity of FHSM, some fast algorithms have been proposed.

In HVDR, 4 neighboring half pixel points in vertical direction and horizontal direction around 'C' illustrated in Fig. 1. are examined to decide the best matching point in each direction. Then, a diagonal point between these two best matching points is also examined. The point having the minimum cost function value among these 5 points and 'C' is decided as a half pixel motion vector. CHPS examines 4 horizontal and vertical half pixel points '2', '4', '5', '7' shown Fig. 1. The point having the minimum cost function value among these 4 points and the point 'C' is decided as a half pixel motion vector.

PPHPS predicts the possible optimal half pixel point by using the cost function values of 5 integer pixel points 'A', 'B', 'C', 'D', 'E' shown Fig. 1. The cost function values of the predicted possible optimal half pixel point and its nearest points are calculated to find the best matching point. The point of the minimum cost function value is decided as a final half pixel MV between this best matching point and the point 'C'.

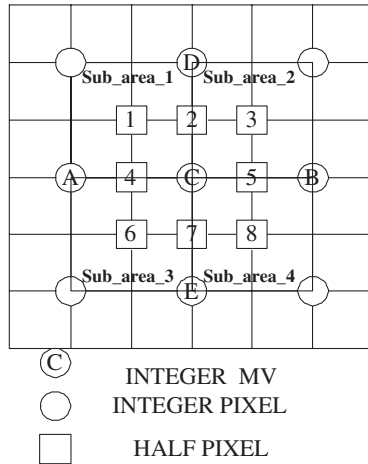


Fig. 1. The Position of integer pixels, half pixels and Subareas

3 The Proposd Method

In order to reduce more computational complexity in half pixel motion estimation, the proposed method exploits spatial correlation among half pixel motion vectors to decide whether half pixel motion estimation is performed or not for the current block. In other words, the proposed method exploits spatially correlated motion vectors depicted in Fig.2. to decide whether the half pixel motion estimation is performed or not for the current block. In case half pixel motion estimation has to be performed to find a half pixel motion vector, we used Yoon’s Fast Half Pixel Search algorithm (YFHPS) as a search algorithm. YFHPS proposed in this paper predicts the possible subarea by using the cost function values of integer pixel points. According to the position of the possible subarea, three half pixel points in its possible subarea are examined to find a half pixel motion vector. At first, YFHPS decides the best horizontal matching point between 2 horizontal integer pixel points ‘A’, ‘B’ depicted in Fig.1. and the best

	MV1_Half (dxh1,dyh1)	MV0_Half (dxh0,dyh0)
MV2_Half (dxh2,dyh2)	Current Block	

MV0_Half (dxh0,dyh0) : half pixel MV of above_right block
MV1_Half (dxh1,dyh1) : half pixel MV of above block
MV2_Half (dxh2,dyh2) : half pixel MV of left block

Fig. 2. Blocks for Spatio-Temporal Correlation Information

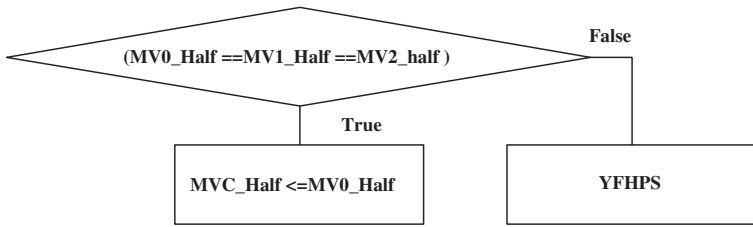


Fig. 3. The Block Diagram of the proposed method

vertical matching point between 2 vertical integer pixel points 'D', 'E' depicted in Fig.1. And then, the possible subarea is selected by using the best horizontal and vertical matching points. According to the position of the possible subarea, three half pixel points in its possible subarea are examined. Finally, the point having the minimum cost function value among these three half pixel points and the point 'C' Fig. 1. is decided as a half pixel motion vector. The block diagram of the proposed method appears in Fig.3. The proposed method is summarized as follows.

- Step 1:** If MV1_Half (dxh1, dyh1) and MV2_Half (dxh2, dyh2) are equal to MV0_Half (dxh0, dyh0), go to Step 2. Otherwise, go to Step 3.
- Step 2:** MV0_Half (dxh0, dyh0) is decided as the half pixel MV of the current block.
- Step 3:** YFHPS is performed to find a half pixel motion vector.

4 Simulation Result

In this section, we show experimental results for the proposed method. The proposed method has been evaluated in the H.263 encoder. Nine QCIF test sequences are used for the experiment. The mean square error (MSE) distortion function is used as the block distortion measure (BDM). The quality of the predicted image is measured by the peak signal to noise ratio (PSNR), which is defined by

$$MSE = \left(\frac{1}{MN} \right) \sum_{m=1}^M \sum_{n=1}^N [x(m, n) - \hat{x}(m, n)]^2 \tag{1}$$

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \tag{2}$$

In Eq. (1), $x(m, n)$ denotes the original image and $\hat{x}(m, n)$ denotes the motion compensated prediction image. For integer pixel motion estimation, Full Search algorithm is adopted. For half pixel motion estimation, we compared FHSM, HVDR, CHPS, PPHPS and YFHPS to the proposed method in both of image quality and search speed. The simulation results in Table 1 and 2 show that the

Table 1. Average PSNR for half pixel motion estimation algorithms

Integer-pel ME method	Full search					
Half-pel ME method	FHSM	HVDR	CHPS	PPHPS	YFHPS	Proposed
Akiyo	34.5	34.41	34.46	34.43	34.5	34.41
Claire	35.05	35.02	35.03	35.05	35.05	35.02
Foreman	29.54	29.52	29.50	29.51	29.51	29.48
M&D	31.54	31.50	31.54	31.52	31.54	31.48
News	30.59	30.49	30.54	30.57	30.57	30.53
Salesman	32.7	32.64	32.67	32.70	32.70	32.66
Silent	31.81	31.80	31.76	31.79	31.80	31.76
Stefan	23.89	23.85	23.86	23.87	23.87	23.83
Suzie	32.19	32.17	32.15	32.19	32.19	32.18

Table 2. The Number of Search points per half pixel MV

	FHSM	HVDR	CHPS	PPHPS	YFHPS	Proposed
Akiyo	8	5	4	3	3	0.3
Claire	8	5	4	3	3	0.9
Foreman	8	5	4	3	3	1.98
M & D	8	5	4	3	3	1.05
News	8	5	4	3	3	0.75
Salesman	8	5	4	3	3	0.48
Silent	8	5	4	3	3	0.96
Stefan	8	5	4	3	3	1.8
Suzie	8	5	4	3	3	1.86

search speed of the proposed method is faster than the other methods (FHSM, HVDR, CHPS, PPHPS and YFHPS) while its PSNR is similar to them except for FHSM. In other words, the proposed method can achieve the search point reduction up to 97% with only 0.01 ~ 0.06 (dB) degradation of image quality. When compared to FHSM.

5 Conclusion

Based on the spatial correlation among half pixel MVs, a simplified method for half pixel motion estimation is proposed in this paper. According to spatially correlated information, the proposed method decides whether half pixel motion estimation is performed or not for the current block. As a result, the proposed method could reduce the computational complexity significantly. Experimental results show that the speedup improvement of the proposed method over FHSM can be up to 4 ~ 26 times faster with a little degradation of the image quality.

References

1. Tham, J.Y., Ranganath, S., Kassim, A.A.: A Novel Unrestricted Center-Biased Diamond Search Algorithm for Block Motion Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*. **8(4)** (1998) 369–375
2. Shan, Z., Kai-kuang, M.: A New Diamond Search Algorithm for Fast block Matching Motion Estimation. *IEEE Transactions on Image Processing*. **9(2)** (2000) 287–290
3. Koga, T., Iinuma, K., Hirano, Y., Iijim, Y., Ishiguro, T.: Motion compensated interframe coding for video conference. In *Proc. NTC81*. (1981) C9.6.1–9.6.5
4. Renxiang, L., Bing, Z., Liou, M.L.: A New Three Step Search Algorithm for Block Motion Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*. **4(4)** (1994) 438–442
5. Lai-Man, P., Wing-Chung, M.: A Novel Four-Step Search Algorithm for Fast Block Motion Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*. **6(3)** (1996) 313–317
6. Yuk-Ying, C., Neil, W.B.: Fast search block-matching motion estimation algorithm using FPGA. *Visual Communication and Image Processing 2000. Proc. SPIE*. **4067** (2000) 913–922
7. Jain, J., Jain, A.: Dispalcement measurement and its application in interframe image coding. *IEEE Transactions on Communications*. **COM-29** (1981) 1799–1808
8. Zhu, C., Lin, X., Chau, L.P.: Hexagon based Search Pattern for Fast Block Motion Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*. **12(5)** (2002) 349–355
9. Ma, K.K., Hosur, P.I.: Report on Performance of Fast Motion using Motion Vector Field Adaptive Search Technique. *ISO/IEC/JTC1/SC29/WG11.M5453* (1999)
10. Tourapis, A.M., Liou, M.L.: Fast Block Matching Motion Estimation using Predictive Motion Vector Field Adaptive Search Technique. *ISO/IEC/JTC1/SC29/WG11.M5866* (2000)
11. Lee, K.H., Choi, J.H., Lee, B.K., Kim, D.G.: Fast two step half pixel accuracy motion vector prediction. *Electronics Letters* **36(7)**(2000) 625–627
12. Cheng, D., Yun, H., Junli, Z.: A Prabolic Prediction-Based, Fast Half Pixel Serch Algorithm for Very Low Bit-Rate Moving Picture Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. **13(6)** (2003) 514–518
13. Cheng, D., Yun, H.: A Comparative Study of Motion Estimation for Low Bit Rate Video Coding. *SPIE* **4067(3)**(2000) 1239–1249

A New Tracking Mechanism for Semi-automatic Video Object Segmentation

Zhi Liu, Jie Yang, and Ningsong Peng

Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University,
No.1954 Huashan Road, Shanghai 200030, People's Republic of China
liuzhi@sjtu.edu.cn

Abstract. This paper proposes a new tracking mechanism for semi-automatic video object segmentation. An interactive video object segmentation tool is presented for the user to easily define the desired video objects in the first frame, and then the video objects are automatically segmented using the proposed tracking mechanism of bi-directional projection. Forward projection is first exploited to locate the current video object with rough boundary information. Watershed segmentation is then applied to the simplified gradient image of the current frame to obtain a reasonable partition. An improved backward projection, which incorporates pixel classification with region classification, is finally performed on some segmented regions in a rather small search range, and the tracking performance is enhanced in respect of both reliability and efficiency. Experimental results for various types of the MPEG-4 test sequences demonstrate an efficient and faithful segmentation performance of the proposed approach.

1 Introduction

As an important issue for the implementation of many content-based multimedia applications supported by MPEG-4, video object segmentation remains a challenging research topic until now. At present, efficient algorithms for automatic video object segmentation only apply to moving objects or some kind of objects with *a priori* knowledge. Therefore, a more practical solution, the so-called semi-automatic video object segmentation [1-10], draws more and more attention in recent years. A typical paradigm of semi-automatic video object segmentation consists of two steps: the user defines the interested video object in the first frame, and then the defined video object is automatically tracked in the rest frames of the sequence.

The first step is extremely important in any semi-automatic video object segmentation algorithms, because the accuracy of the segmented video objects directly determines the success or failure of the following tracking process. A user-friendly segmentation tool should be provided for the user to conveniently define the video objects, and user interaction activity should be minimized to improve the segmentation efficiency. However, the flexibility and efficiency of user interaction are rarely considered as important as the algorithm itself in most

existing approaches. The most common way of user interaction is to delineate an approximate contour clinging to the video object [1,2]. However, it is a burdened job to move mouse along the true object contour, especially when the shape of the object is complex. For those approaches associated with snake model, a considerable number of control points around the object contour need to be selected one by one [3,4]. Region selection is a more natural way to define a video object, but an excessive number of regions still need to be selected at different partition levels [5]. In this paper, we present an efficient and flexible video object segmentation tool [6] to define the interested video object in the first frame, which takes both advantages of two user interaction ways, i.e. marker drawing and region selection.

For the second step, many existing approaches adopt a two-step configuration to track the video object [1,2,7,8], i.e., first project the previous object to the current frame using some kind of parametric motion model, and then refine the projected object boundary. The underlying tracking mechanism is forward projection, which works well for rigid objects with translation motion. For non-rigid objects with multiple motions, irregular boundaries and uncertain holes may appear on the video objects, and inevitable post-processing is needed for boundary refinement. In contrast with forward projection, backward projection [9,10] is suitable to deal with non-rigid objects, and needs no further refinements. Each segmented region in the current frame is projected to the previous frame, and then it is assigned to the current video object if the majority of the projected region overlaps the previous video object. In nature, it is a region classification approach rather than a tracking approach. However, it is not an efficient way to backward project all segmented regions for classification. Another problem may occur when a segmented region overlaps the video object and the background, which causes peninsulas or gaps to appear on the video object no matter what classification is assigned to. In this paper, we propose a new tracking mechanism of bi-directional projection, which is more efficient due to the combination of backward projection and forward projection, and ensures the visual quality of the tracked video objects by incorporating pixel classification with region classification.

The rest of the paper is organized as follows. Section 2 describes the proposed tracking mechanism of bi-directional projection for semi-automatic video object segmentation. Experimental results for different types of the MPEG-4 test sequences are shown in section 3, and conclusions are given in section 4.

2 The Proposed Tracking Mechanism of Bi-directional Projection

In this section, we propose a new tracking mechanism for semi-automatic video object segmentation. The whole process of video object segmentation consists of five steps: *user interaction*, *forward projection*, *spatial segmentation*, *backward projection* and *post-processing*. In the first step, an interactive video object segmentation tool is provided for the user to define the interested video ob-

jects. From the second to the fourth step, the proposed tracking mechanism of bi-directional projection is exploited to segment the video objects in the subsequent frames. Some post-processing work that is necessary for the whole tracking process is performed in the last step.

2.1 User Interaction

In order to facilitate the user to easily extract the desired video object, we combine two ways of user interaction, i.e., marker drawing and region selection. The whole procedure of interactive video object segmentation consists of three steps: *marker drawing*, *automatic video object extraction*, and *user correction*. A screen shot of our graphical user interface (GUI) is shown in Fig. 1, which is exploited to clearly describe each step in the following.

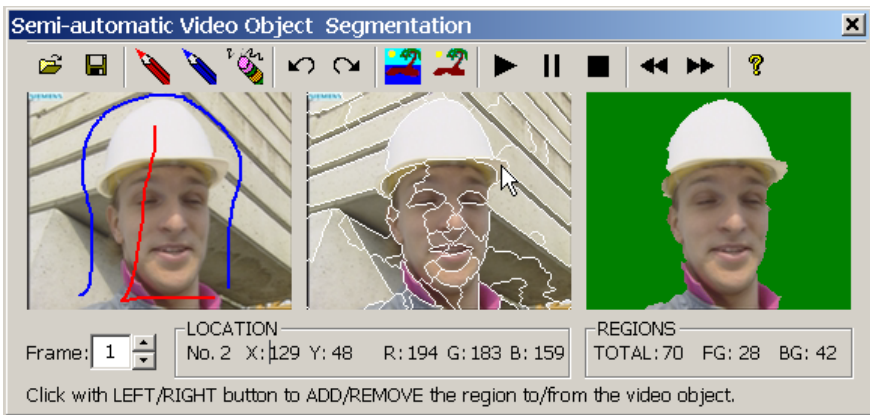


Fig. 1. A screen shot of our graphical user interface

First, the user draws scribbles of different colors to roughly mark the video object and the background. As shown in Fig. 1, a red scribble marks the interested video object, and a blue scribble marks the background in the left window of the GUI. Scribble drawing is more convenient and flexible for the user to experience. It usually takes a few seconds, which is faster than the way of contour drawing and control points selection. Then, the interested video object is automatically extracted using a fast seed region merging approach proposed in [6]. This step does not need any user interference. The spatial segmentation result and the automatically extracted video object are shown in the middle and the right window, respectively. If not satisfied with the automatically extracted video object, the user can make correction. In our GUI, the user is prompted to click with the left button on the region to add it to the video object, while click with the right button to remove. The number of mouse clicks depends on the image content and the marker drawing, usually less than 2 in our many experiments. For example, the left image in Fig. 1 shows low contrast between the

helmet of “foreman” and the background, and the region selected by the mouse in the middle image is merged into the background in the process of seeded region merging. The user just needs to click once with the left button to obtain the desired video object shown in the right image.

2.2 Forward Projection

The objective of forward projection is to locate the video object with rough boundary information, which is derived from the motion estimation. For each contour pixel $I_{n-1}(x, y)$ of the previous video object vo_{n-1} (see Fig. 2(a)), the motion vector $(u(x, y), v(x, y))$ is estimated using the three-step searching method [11] to minimize the following prediction error

$$e(x, y) = \min_{u,v} \sum_{i=-N}^N \sum_{j=-N}^N \|I_{n-1}(x+i, y+j) - I_n(x+u+i, y+v+j)\| \quad (1)$$

The size of the matching block equals $(2N + 1) \times (2N + 1)$. In our experiments, N is set to 2, and the search range for the motion vector $(u(x, y), v(x, y))$ is set to $[-7, 7]$ for all sequences. Forward projection is performed on all contour pixels of vo_{n-1} , denoted by a pixel set ct_{n-1} . The projection of ct_{n-1} in the current frame I_n can be denoted by another pixel set p_n (see the black pixels in Fig. 2(b)).

$$p_n = \{(x + u(x, y), y + v(x, y)) | (x, y) \in ct_{n-1}\} \quad (2)$$

These projected pixels in p_n may not exactly fall onto the true contour ct_n of the video object vo_n in the current frame, and they generally cannot form a closed contour. All pixels in p_n are then dilated with a disk-shaped structuring element E_d to obtain a band area B_n (see Fig. 2(b)) to accommodate the rotation, scale change and deformation of the video object. The radius of E_d is set to 15 by the experiment, which is enough for most video sequences to ensure that the true contour ct_n locates in B_n . The approximate translation vector (T_{n-1}^u, T_{n-1}^v) for the video object is estimated using the average of motion vectors for all the pixels in ct_{n-1}

$$T_{n-1}^u = \frac{\sum_{(x,y) \in ct_{n-1}} u(x, y)}{|ct_{n-1}|}, T_{n-1}^v = \frac{\sum_{(x,y) \in ct_{n-1}} v(x, y)}{|ct_{n-1}|} \quad (3)$$

This vector (T_{n-1}^u, T_{n-1}^v) reflects a global translation movement of the video object if an apparent translation exists, which will be used in the subsection *backward projection*.

2.3 Spatial Segmentation

The watershed segmentation algorithm [12] is exploited to partition the current frame I_n into a set of regions (see Fig. 2(c)). In fact, only the band area B_n

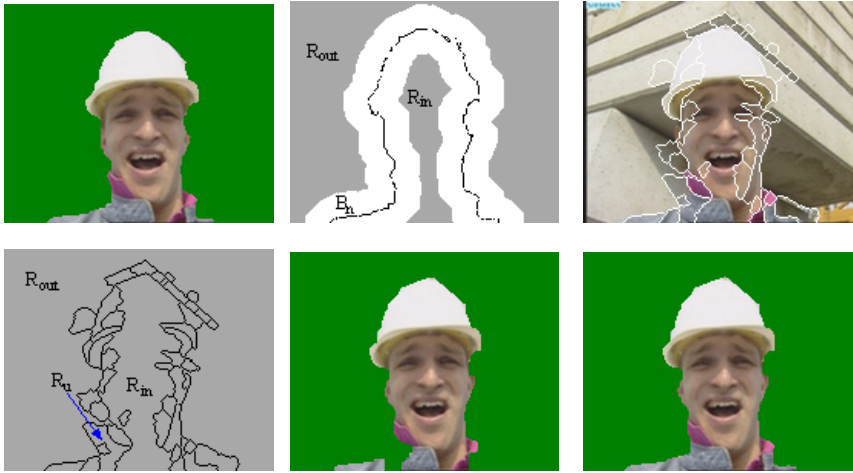


Fig. 2. A pictorial description of the proposed tracking mechanism of bi-directional projection (top row: (a)-(c) bottom row: (d)-(f))

needs to be partitioned into some regions, which need to be backward projected to determine whether they belong to the current video object or not. The area R_{in} inside B_n definitely belongs to the current video object, while the area R_{out} outside B_n belongs to the background (see Fig. 2(b)). Therefore, only the pixels in B_n require gradient calculation, while the gradient of all other pixels is merely set to zero. It also saves the computation time of watershed segmentation, because the areas that are not covered by B_n can be simply flooded as the lowest flat catchment basins. The following gives a detailed description of spatial segmentation itself.

The gradient image g of the color image f in YUV space is estimated by the delicate method proposed by Di Zenzo [13]. However, the noise in the gradient image causes the inevitable problem of over-segmentation when directly applying watershed segmentation. In order to obtain a moderate segmentation result, we propose a simplification step to remove insignificant local minima in the gradient image. First, g is dilated with a 3×3 cross-shaped structuring element E , and the dilated image is elevated by a height h to get the marker image, $g_m = (g \oplus E) + h$. Then, the reconstruction of g from g_m by geodesic erosion [14] is performed to obtain the simplified gradient image, $g_s = \varphi^{(rec)}(g_m, g)$. Now we can apply the watershed segmentation algorithm to the image g_s to obtain the label image f_l , which shows a reasonable partition of the original image f . The only one parameter h is set to 2 by the experiment, which is the second smallest value that leads to a reasonably finer partition (see Fig. 2(c)).

2.4 Backward Projection

The objective of backward projection is to find the true contour ct_n of the current video object vo_n in B_n . The segmented regions, excluding the regions R_{in} and

R_{out} (see Fig. 2(d)), are backward projected to determine their classifications. For each region R_i , the backward motion vector (u_i, v_i) is estimated to minimize the following prediction error

$$e_i = \min_{u_i, v_i} \sum_{(x, y) \in R_i} \|I_n(x, y) - I_{n-1}(x + u_i, y + v_i)\| \tag{4}$$

A small search range $[-T_{n-1}^u - 3, -T_{n-1}^v + 3]$ is set for (u_i, v_i) , because the vector (T_{n-1}^u, T_{n-1}^v) has already reflected a possible apparent translation of the video object. The backward projected region R'_i in the previous frame I_{n-1} can be denoted by the following formula

$$R'_i = \cup_{(x, y) \in R_i} (x + u_i, y + v_i) \tag{5}$$

The classification of R_i can be determined from the intersecting area of R'_i and vo_{n-1} . A natural method [9] is that R_i is classified into vo_n if the majority of R'_i intersects with vo_{n-1} , otherwise classified into the background. However, it is not a robust method to always guarantee the visual quality of the segmented video objects during the tracking process. Specifically, binary classification is not suitable for such a segmented region that overlaps the video object and the background at the same time. If such a region (see the region R_u in Fig. 2(d)) is classified into the video object, a peninsula appears on the video object; otherwise a gap appears (see Fig. 2(e)). In order to deal with such a problem, we propose a robust approach to improve the method in [9].

The ratio of the intersecting area of R'_i and vo_{n-1} to the area of R'_i is defined by the following formula

$$\theta_i = \frac{A[R'_i \cap vo_{n-1}]}{A[R'_i]} \tag{6}$$

where $A[.]$ denotes the area operation. The value of θ_i indicates three different types of region, that is, a fairly higher value shows the region R_i belongs to the video object, a fairly lower value shows R_i is a part of the background, and a moderate value shows R_i may overlaps the video object and the background at the same time. For the first and the second cases, the whole region is assigned to the video object or the background based on the following criterion

$$\begin{cases} R_i \in vo_n, \text{if } \theta_i > T_h \\ R_i \notin vo_n, \text{if } \theta_i < T_l \end{cases} \tag{7}$$

For the third case, $T_l \leq \theta_i \leq T_h$, pixel classification in the region R_i is performed using the following criterion

$$\begin{cases} (x, y) \in vo_n, \text{if } (x + u_i, y + v_i) \in vo_{n-1} \\ (x, y) \notin vo_n, \text{if } (x + u_i, y + v_i) \notin vo_{n-1} \end{cases} \tag{8}$$

Since the value of θ_i lies in the range of $[0, 1]$, T_h should be greater than 0.5, i.e., $T_h = 0.5 + \Delta (\Delta > 0)$. The other parameter T_l is set to the margin value

Δ . Therefore, both pixel classification and region classification hold a half of the whole range. In our experiments, the margin value Δ is set to 0.15 for all test sequences, and these two criteria lead to a reliable tracking performance (see Fig. 2(f)).

2.5 Post-processing

In the previous subsection, only the regions are considered in the process of backward projection, while those boundaries (watershed lines) between different regions are not classified. Therefore, a closing morphological operation is first performed to fill the watershed lines in the video object. It is the closed video object that is propagated in the tracking process. Since a closing operation is needed, an opening operation is also performed subsequently. Here, a cascade of closing and opening operation also smoothes the boundary of the video object, which sometimes enhances the visual quality of the segmented video object. The structuring element for both morphological operations is a 5×5 square, which can achieve a good tradeoff between the accuracy and the smoothness of video object boundaries.

3 Experimental Results

We use several MPEG-4 test sequences to test the proposed tracking mechanism of bi-directional projection, and the experimental results for three of them are shown Figs. 3–5. These sequences represent different levels of spatial detail and movement in real situations. The first sequence *Mother and Daughter* (176×144) is a MPEG-4 class A sequence, with low spatial detail and low amount of movement. The background is uniform and static, and the motion of human bodies is relatively small. The second sequence *Foreman* (176×144) is a MPEG-4 class B sequence, with medium spatial detail and low amount of movement. The background is complex and shows low contrast with the talking person, and the camera motion is also apparent besides the non-rigid motion of the person. The third sequence *Table Tennis* (352×240) is a MPEG-4 class C sequence, with high spatial detail and medium amount of movement. Several moving objects appear on the clutter background. The interested video object is the arm holding the racket, which mixes different rigid motions of the arm, the hand and the racket.

For all sequences, the initial video objects can be easily obtained using our interactive video object segmentation tool, and an example for the sequence *Foreman* is described in subsection 2.1. In our experiments, it takes about 5 seconds to obtain the desired video object shown in the first image of each figure. The tracking results are shown in the latter three images in turn. For all sequences, the accurate and reliable video objects are obtained using the proposed tracking mechanism of bi-directional projection.

These experiments are performed on a low-end AMD Athlon 1.53GHz PC. The average processing time per frame using our bi-directional projection approach and the backward projection approach in [9] is shown in Table 1. The



Fig. 3. Experimental results for *Mother and Daughter* (Frame. 1, 20, 60, 100)



Fig. 4. Experimental results for *Foreman* (Frame. 1, 20, 60, 100)



Fig. 5. Experimental results for *Table Tennis* (Frame. 1, 15, 25, 40)

Table 1. Average processing time for a frame of the three sequences (msec)

Test sequences	Bi-directional projection				Backward projection		
	<i>FP</i>	<i>Seg.</i>	<i>BP</i>	<i>Total</i>	<i>Seg.</i>	<i>BP</i>	<i>Total</i>
<i>Mother and Daughter</i>	47	37	81	165	43	281	324
<i>Foreman</i>	43	42	68	153	58	275	333
<i>Table tennis</i>	95	131	181	407	154	981	1135

same values are set to the related parameters in both approaches. Compared with the approach in [9], our approach needs to consume some time on forward projection, but sharply reduce the time on backward projection, and spatial segmentation to some extent. For the three sequences, the total processing time of our approach is 51%, 46%, and 36% of the approach in [9], which demonstrate the improved segmentation efficiency of our approach. It is promising that more efficiency can be gained after code optimization or using a higher speed processor.

4 Conclusions

Video object segmentation is an inevitable necessity for MPEG-4 related multimedia applications. In this paper, we propose a new tracking mechanism for

semi-automatic video object segmentation. An interactive video object segmentation tool is presented for the user to easily define the video objects. The extracted video objects are then automatically segmented using the proposed tracking mechanism of bi-directional projection, which extends backward projection with the combination of forward projection. The proposed tracking mechanism produces more reliable video objects for different types of video sequences, and improves the segmentation efficiency by a factor of two. In our future work, we will consider to use some high level features of the video object in the tracking process to further improve the segmentation reliability and accuracy.

References

1. Kim, M., Jeon, J.G., Kwak, J.S., Lee, M.H., Ahn, C.: Moving Object Segmentation in Video Sequence by User Interaction and Automatic Object Tracking. *Image and Vision Computing*. 5 (2001) 245–260
2. Guo, J., Kim, J.W., Kuo, C.-C.J.: An Interactive Object Segmentation System for MPEG Video. *IEEE Int. Conf. Image Processing*, Vol. 2 (1999) 140–144
3. Luo, H.T., Eleftheriadis, A.: An Interactive Authoring System for Video Object Segmentation and Annotation. *Signal Processing: Image Communication*. 7 (2002) 559–572
4. Sun, S.J., Haynor, D.R., Kim, Y.M.: Semiautomatic Video Object Segmentation using Vsnakes. *IEEE Trans. Circuits Syst. Video Technol.* 1 (2003) 75–82
5. Cooray, S., O'Connor, N., Marlow, S., Murphy, N., Curran, T.: Hierarchical Semi-automatic Video Object Segmentation for Multimedia Applications. *Proc. SPIE Internet Multimedia Management Systems II*, Vol. 4519 (2001) 10–19
6. Zhi L., Jie Y.: Interactive Video Object Segmentation: Fast Seeded Region Merging Approach. *Electronics Letters*. 5 (2004) 302–304
7. Gu, C., Lee, M.C.: Semiautomatic Segmentation and Tracking of Semantic Video Objects. *IEEE Trans. Circuits Syst. Video Technol.* 5 (1998) 572–584
8. Lim, J., Cho, H.K., Beom Ra, J.: An Improved Video Object Tracking Algorithm Based on Motion Re-estimation. *IEEE Int. Conf. Image Processing*, Vol.1 (2000) 339–342
9. Gu, C., Lee, M.C.: Semantic Video Object Tracking Using Region-based Classification. *IEEE Int. Conf. Image Processing*, Vol. 3 (1998) 643–647
10. Gatica-Perez, D., Sun, M.T., Gu, C.: Semantic Video Object Extraction Based on Backward Tracking of Multivalued Watershed. *IEEE Int. Conf. Image Processing*, Vol. 2 (1999) 145–149
11. Tekalp, A.M.: *Digital Video Processing*. Tsinghua University Press, Beijing (1998)
12. Vincent, L., Soille, P.: Watersheds in Digital Spaces: an Efficient Algorithm Based on Immersion Simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* 6 (1991) 583–598
13. Di Zenzo, S.: A Note on the Gradient of a Multi-image. *Computer Vision Graphics Image Processing*. 1 (1986) 116–125
14. Vincent, L.: Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient algorithms. *IEEE Trans. Image Processing*. 2 (1993) 176–201

A Visual Model for Estimating the Perceptual Redundancy Inherent in Color Images

Chun-Hsien Chou and Kuo-Cheng Liu

Department of Electrical Engineering,
Tatung University, Taiwan
chou@ttu.edu.tw, kcliu@tth.edu.tw
<http://www.ttu.edu.tw>

Abstract. An efficient compression algorithm that is transparent to human visual perception is highly expected in representing high-quality color image as the resource of storage or transmission bandwidth is limited. Since the human eyes are not perfect sensors for discriminating color signals of small differences, there exist spaces of perceptual redundancy in color images. This paper presents a visual model for estimating perceptual redundancies of color images in terms of a triple of values, each for a color channel as a visibility threshold of distortion, or just noticeable difference (JND). The visual model is built on defining a perceptually indistinguishable region for each color in a color space by mapping colors, which are barely distinguishable from the target color in a perceptually uniform color space (PUCS), to the target color space. To justify the proposed color visual model, the estimated perceptual redundancy is utilized to improve the performance of JPEG-LS. Simulation results show that the perceptual performance in terms of the inspected visual quality and the amount of perceivable errors of the perceptually tuned JPEG-LS coder is superior to that of the un-tuned coder.

1 Introduction

The exploitation of the perceptual redundancy inherent in digital images has been proven successful in enhancing the performance of many digital image processing systems such as those for data compression and watermarking [1-4]. However, most of these systems and related researches focused only on the estimation and exploitation of the perceptual redundancy inherent in luminance signals. It is expected that there exists a considerable amount of perceptual redundancy in chrominance signals, but lacks for a systematic approach to quantitatively measure the chromatic perceptual redundancy.

With colorimetric coordinates for specifying each color in a color space, the perceptual redundancy of a color pixel can be quantitatively measured by calculating the distance between the target color and each of the colors that are barely different from the target color. However, colors in many color spaces, such as *RGB*, *XYZ*, *YUV*, and *YCbCr* are not uniformly distributed in the sense that Euclidean distances between colors are not closely correlated with color differences perceived by humans. That is, the colors that are indiscernible from the

target color form a perceptually indistinguishable region with irregular shape in these color spaces. The perceptually indistinguishable region always provides large amount of perceptual redundancy which can be used to improve the performance of image processing applications. It also explains why a color visual model should be investigated to conveniently define the perceptually indistinguishable region for quantifying the perceptual redundancy of colors.

The spaces *CIELAB* and *CIELUV* recommended by International Commission on Illumination (*CIE*) in 1976 were such perceptually uniform color spaces through nonlinear transformations of tristimulus *XYZ* as to overcome the non-uniformity of color spaces that had been discussed by MacAdam and others from the early 1940s [5]. Even though *CIELAB* is not a perfect uniform color space, the color distance between two colors is considerably correlated with the perceptual difference. If a color space is perfectly uniform in color metric, the locus of colors which are not perceptually different from a given color forms a sphere with a radius equal to the so-called *just noticeable color difference* (JNCD) [6]. By exploiting the sphere of JNCD in the uniform color space and the transformation of color spaces, the perceptually indistinguishable region of the color in target color spaces can be conveniently defined.

In this paper, a color visual model is proposed to estimate the perceptual redundancy inherent in color images. The proposed visual model employs the Euclidean distance equation to find color samples that are located on the surface of the JNCD sphere of a target color. Under a specified viewing condition, the relationship between the numerical color difference and the perceived color difference will be investigated and the threshold of just perceptible difference will be defined. With the perceptibility threshold of color difference, an algorithm for locating sample colors that are close to the surface of the JNCD sphere around a target color is developed. The found color samples are then transformed to the target color space for defining the perceptually indistinguishable region and quantifying the perceptual redundancy in each color channel. To justify the proposed color visual model, the perceptual performance between a JPEG-LS coder and its perceptually tuned counterpart is compared.

2 The Proposed Human Visual Model

2.1 Human Visual Model

The proposed visual model for estimating perceptual redundancies of color images provides JND values for each color pixel in spatial domain. The problem that arises when quantifying the perceptual redundancy in terms of visibility threshold of distortion is that the perceptually indistinguishable region for each color in the target color space needs to be defined to ensure the visibility threshold of each channel. In our model, we develop an algorithm for finding color samples that are located close to the surface of the JNCD sphere around the target color in the *CIELAB* and establish a criterion for defining a perceptually indistinguishable region around the target color in the working color space from those sampled JNCD colors transformed from the *CIELAB*. Since the luminance

contrast sensitivity function (CSF) is significantly higher than the chromatic CSFs, the sampling algorithm is constrained in the sense that the luminance differences between sample colors and the target color should be smaller than the perceptibility threshold of luminance component. It is believed that the color samples with corresponding color differences set to be luminant JND of target color can be used to approximate the perceptually indistinguishable region. The demonstration of the proposed human visual model is shown in Fig. 1.

As shown in Fig. 1(a), the perceptual redundancy of the color \mathbf{k} in the $YCbCr$ color image is to be estimated. Let (Y_k, Cb_k, Cr_k) be the tristimulus values of color \mathbf{k} . The JND value of the luminance signal is obtained by means of the same perceptual model conducted in [1] and expressed by JND_Y . To develop the color sampling algorithm in the $CIE\ L A B$, target color The proposed visual model for estimating perceptual redundancies of color images provides JND values for each color pixel in spatial domain. The problem that arises when quantifying the perceptual redundancy in terms of visibility threshold of distortion is that the perceptually indistinguishable region for each color in the target color space needs to be defined to ensure the visibility threshold of each channel. In our model, we develop an algorithm for finding color samples that are located close to the surface of the JNCD sphere around the target color in the $CIE\ L A B$ and establish a criterion for defining a perceptually indistinguishable region around the target color in the working color space from those sampled JNCD colors transformed from the $CIE\ L A B$. Since the luminance contrast sensitivity function (CSF) is significantly higher than the chromatic CSFs, the sampling algorithm is constrained in the sense that the luminance differences between sample colors and the target color should be smaller than the perceptibility threshold of luminance component. It is believed that the color samples with corresponding color differences set to be luminant JND of target color can be used to closely map to the perceptually indistinguishable region. The demonstration of the proposed human visual model is shown in Fig. 1.

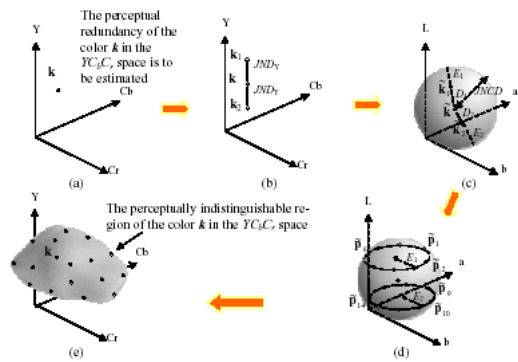


Fig. 1. The demonstration of the proposed color visual model.

As shown in Fig. 1(a), the perceptual redundancy of the color \mathbf{k} in the $YCbCr$ color image is to be estimated. Let $(Y_{\mathbf{k}}, Cb_{\mathbf{k}}, Cr_{\mathbf{k}})$ be the tristimulus values of color \mathbf{k} . The JND value of the luminance signal is obtained by means of the same perceptual model conducted in [1] and expressed by JND_Y . To develop the color sampling algorithm in the *CIELAB*, target color \mathbf{k} and the JND_Y contaminated colors \mathbf{k}_1 and \mathbf{k}_2 should be transformed to *CIELAB*. The transformed colors are given by $\tilde{\mathbf{k}}, \tilde{\mathbf{k}}_1$ and $\tilde{\mathbf{k}}_2$, respectively and depicted in Fig. 1(b), (c). The JND_Y threshold of color pixel \mathbf{k} is nonlinearly mapped into *CIELAB* and presented by D_1 and D_2 .

In *CIELAB*, the colors spread on the surface of JNCD sphere around the target color are expressed by a color set.

$$\mathbf{S}_{\tilde{\mathbf{k}}} = \{\tilde{\mathbf{s}}_i \mid \|\tilde{\mathbf{s}}_i - \tilde{\mathbf{k}}\| = JNCD\} \tag{1}$$

where $\|\cdot\|$ means the Euclidean distance. To find color samples located on this surface and satisfy the requirement for closely defining the perceptually indistinguishable region of the target color \mathbf{k} from those transformed color samples, we address on two color subsets of $\mathbf{S}_{\tilde{\mathbf{k}}}$ in which colors contaminated with luminant detection threshold are considered.

$$\mathbf{C}_{\tilde{\mathbf{k}}_1} = \{\tilde{\mathbf{c}}_m \mid \tilde{\mathbf{c}}_m \in \mathbf{S}_{\tilde{\mathbf{k}}}, |L_{\tilde{\mathbf{c}}_m} - L_{\tilde{\mathbf{k}}}| = |L_{\tilde{\mathbf{k}}_1} - L_{\tilde{\mathbf{k}}}| \} \tag{2}$$

$$\mathbf{C}_{\tilde{\mathbf{k}}_2} = \{\tilde{\mathbf{c}}_n \mid \tilde{\mathbf{c}}_n \in \mathbf{S}_{\tilde{\mathbf{k}}}, |L_{\tilde{\mathbf{k}}} - L_{\tilde{\mathbf{c}}_n}| = |L_{\tilde{\mathbf{k}}} - L_{\tilde{\mathbf{k}}_2}| \} \tag{3}$$

In the proposed model, there are 16 color samples are chosen from these two subsets for color mapping.

$$\{\tilde{\mathbf{p}}_m\}_{m=1\dots 8}, \tilde{\mathbf{p}}_m \in \mathbf{C}_{\tilde{\mathbf{k}}_1} \tag{4}$$

$$\{\tilde{\mathbf{p}}_n\}_{n=1\dots 8}, \tilde{\mathbf{p}}_n \in \mathbf{C}_{\tilde{\mathbf{k}}_2} \tag{5}$$

As shown in Fig. 5(d), these color samples around reference color $\tilde{\mathbf{k}}_1$ and $\tilde{\mathbf{k}}_2$ are spreading on the planes paralleled AB diagram in *CIELAB*.

By utilize the linear property provided by the *CIELAB* color space, the additive and subtractive operations can be directly applied to color difference measurement for human perception. Since the values of D_1 and D_2 are the perceptual redundancy contributed by luminance component, the values of E_1 and E_2 evaluated by the difference between $JNCD$ and D_1 and D_2 , respectively, can be reasonably regarded as the redundancy contributed by chrominance component appear in *CIELAB*. With the value E_1 and E_2 , the color samples can be simply found as follows for reference color $\tilde{\mathbf{k}}_1$ and $\tilde{\mathbf{k}}_2$, respectively.

$$(L_{\tilde{\mathbf{p}}_m}, a_{\tilde{\mathbf{p}}_m}, b_{\tilde{\mathbf{p}}_m}) = (L_{\tilde{\mathbf{k}}_1}, a_{\tilde{\mathbf{k}}_1} + \Delta a_{\tilde{\mathbf{k}}_1}, b_{\tilde{\mathbf{k}}_1} + \Delta b_{\tilde{\mathbf{k}}_1}) \tag{6}$$

$$(L_{\tilde{\mathbf{p}}_n}, a_{\tilde{\mathbf{p}}_n}, b_{\tilde{\mathbf{p}}_n}) = (L_{\tilde{\mathbf{k}}_2}, a_{\tilde{\mathbf{k}}_2} + \Delta a_{\tilde{\mathbf{k}}_2}, b_{\tilde{\mathbf{k}}_2} + \Delta b_{\tilde{\mathbf{k}}_2}) \tag{7}$$

where $\Delta a_{\tilde{\mathbf{k}}_1} = Re(E_1 e^{jk\theta})$, $\Delta b_{\tilde{\mathbf{k}}_1} = Im(E_1 e^{jk\theta})$, $\Delta a_{\tilde{\mathbf{k}}_2} = Re(E_2 e^{jk\theta})$, $\Delta b_{\tilde{\mathbf{k}}_2} = Im(E_2 e^{jk\theta})$ for $k = 0 \dots 7$ and $\theta = 45^\circ$.

Color samples are transformed back to the $YCbCr$ and given by $\mathbf{p}_1, \mathbf{p}_2, \dots$, and \mathbf{p}_3 , respectively. The perceptually indistinguishable region can be approximately approached by these mapping colors as shown in Fig. 1(e). Finally, the estimated JND value of pixel \mathbf{k} for Cb and Cr channels can be conservatively obtained by taking the minimal distance between color \mathbf{k} and those mapping colors along Cb and Cr channel.

$$JND_{Cb} = \min_{i=1\dots16} |Cb_{\mathbf{p}_i} - Cb_{\mathbf{k}}| \tag{8}$$

$$JND_{Cr} = \min_{i=1\dots16} |Cr_{\mathbf{p}_i} - Cr_{\mathbf{k}}| \tag{9}$$

where $(Y_{\mathbf{k}}, Cb_{\mathbf{k}}, Cr_{\mathbf{k}})$ and $(Y_{\mathbf{p}_i}, Cb_{\mathbf{p}_i}, Cr_{\mathbf{p}_i})$ are tristimulus values of color \mathbf{k} and \mathbf{p}_i in $YCbCr$, respectively.

2.2 Estimation of Adjustable JNCD Sphere in the CIELAB

It is noted that the JNCD threshold ($JNCD = 3$) of each color pixel is much lower in our experiments while taking the local properties of color images into account. Since the luminance CSF is significantly higher than the chromatic CSFs, only luminance dominated properties of the HVS are employed to adjust, from data in spatial domain, the JNCD value associated with each pixel of the color image. Besides, the considerable non-uniformity of *CIELAB* results in the fact that the JNCD threshold will be changed with the chroma of a target color according to the color discrimination data set [7]. Hence, two adjustments are exploited to adjust the JNCD threshold of each color pixel with its related characteristics. In our model, the adjusted JNCD thresholds of different colors are expressed as follows.

$$AJND_{Cr} = JNCD \cdot s_L(Y, \Delta Y) \cdot s_C(a, b) \tag{10}$$

where $s_L(Y, \Delta Y)$ and $s_C(a, b)$ denote the luminance dominant adjustment and the chroma adjustment, respectively.

3 Application to JPEG-LS

Based on the previous discussion, the proposed color visual model can be regarded as an independent module to various coding algorithm. In this paper, we blend it into the JPEG-LS algorithm to enhance the corresponding performance. The block diagram of the perceptual tuned JPEG-LS algorithm is depicted in Fig. 2.

JPEG-LS is an emerging new ISO/ITU standard for lossless and near-lossless image compression [7]. The near-lossless coding mode of JPEG-LS mainly consists of three operational parts including prediction, statistical modeling, and entropy coding. By analyzing the texture within a causal template, the context-based statistical model selects an appropriate mode to code the current pixel. The current pixel is coded in run-length mode if it is located in a smooth area.

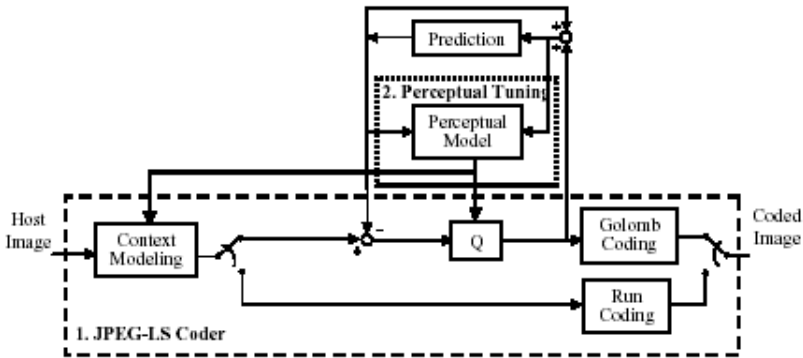


Fig. 2. Block diagram of perceptually tuned JPEG-LS algorithm.

On the other hand, the pixel is coded in regular mode if it is located in a region with edges. In the regular mode, the context determination procedure is followed by a prediction procedure. Based on the same context template, the predictive value is determined by a simple scheme of edge detection among the context pixels. The prediction residual is then quantized by a uniform quantizer of which the quantization step size is set by the information loss parameter *NEAR* such that the maximum distortion will not exceed *NEAR*. Although the coding distortion is bounded by the parameter *NEAR*, the performance of the JPEG-LS coder can be further promoted by shaping the noises such that coding distortion is imperceptible. Therefore, to reach an optimized performance that maintains the highest possible visual quality at the lowest possible bit rates, the occurrence of the run-length coding mode should be maximized under the conditions that the resulting distortion is not perceivable, and the quantization stepsize for coding prediction residual should be perceptually tuned in a way that the distortion introduced by quantization is not perceivable.

In this paper, the near-lossless coding performance of the JPEG-LS coder is improved by replacing the context modeling with a perceptually tuned modeler for determining the coding mode. If the distortion introduced by coding the current pixel in run-length mode exceeds the error visibility threshold as estimated by the proposed color visual model, the current pixel is coded in regular mode. Otherwise, the pixel is coded in run-length mode. The perceptual redundancy estimated by the proposed visual model is also used to adaptively adjust the quantizer stepsize for coding the prediction residue such that quantization distortion is not perceivable in the reconstructed image.

4 Simulation Results

Color images of size 512×512 are used as tested images, where the tristimulus values of each pixel are represented by 24 bits. To show that the model works, a



Fig. 3. (a) Original “Lena” (b) JND contaminated version.

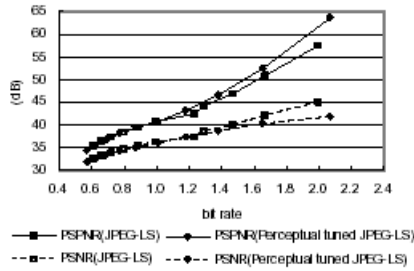


Fig. 4. Plot of PSNR and PSPNR versus bit rate for “Baboon”.

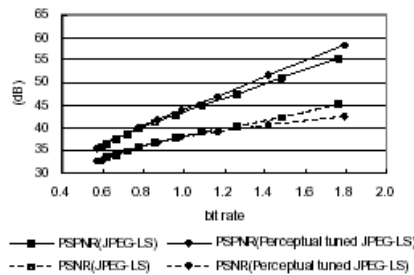


Fig. 5. Plot of PSNR and PSPNR versus bit rate for “Lena”.

subjective test that compares the perceptual quality of the image contaminated by the noises of JND profiles with the original image is shown in Fig. 3. The JND contaminated version with 33.2dB in PSNR has the same perceptual quality as the original image at a viewing distance of about 6 times the image height.

The comparison between perceptual JPEG-LS and JPEG-LS is illustrated by using a fidelity criterion. A fidelity measure, peak signal-to-perceptible-noise ratio (PSPNR), extended from [1] is adopted here. The compression ratio required by the proposed coding scheme can be achieved by adjusting the perceptual

information loss with the *minimally noticeable distortion* (MND).

$$MND(i, j) = JND(i, j) + d \quad (11)$$

where d is the offset factor. The MND profiles of different distortion levels are hence required to optimize the coding efficiency at different bit rate budget.

The plots of the values of PSNR and PSPNR versus the required bit rates for both the proposed algorithm and JPEG-LS algorithm are shown in Fig. 4 and Fig. 5 with “Baboon” and “Lena”, respectively. The visual performance of the perceptual tuned JPEG-LS is better than that of the JPEG-LS algorithm. At the same bit rate, the amount of perceptual redundancy removed by the perceptual JPEG-LS is larger than that by JPEG-LS.

5 Conclusions

In this paper, a visual model for estimating perceptual redundancies of color images is proposed. Through the JNCD in the uniform color space and the transformation of color spaces, the perceptually indistinguishable region of the color in target color spaces is defined and the perceptual redundancy is successfully quantified in terms of JND values for each component. The validity of the proposed visual model is tested by the subjective quality between the original image and the JND contaminated version. The estimated perceptual redundancy is further utilized to improve the performance of JPEG-LS. Our further work is to construct the rate controlled algorithm by applying the visual model to the selection of coding mode for JPEG-LS.

References

1. C. H. Chou and Y. C. Li, “A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 467-476, Dec. 1995.
2. I. Höntsch and L. J. Karam, “Locally adaptive perceptual image coding,” *IEEE Trans. Image Processing*, vol. 9, pp. 1472-1483, Sept. 2000.
3. A. B. Watson, G. Yang, J. A. Solomon, and J. Villasenor, “Visibility of wavelet quantization noise,” *IEEE Tran. Image Processing*, vol. 6, pp. 1164-1175, Aug. 1997.
4. C. H. Chou and T. L. Wu, “Embedding color watermarks in color images,” *EURASIP Journal of Applied Signal Processing*, pp.32-40, vol.2003, no. 1, Jan. 2003.
5. D. L. MacAdam, “Specification of small chromaticity differences,” *J. Opt. Soc. Am.*, vol. 33, pp. 18-26, 1943.
6. M. Mahy, L. Van Eyckden, and A. Oosterlinck, “Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV,” *Color Res. Appl.*, vol. 19, pp. 105-121, Apr. 1994.
7. ISO/IEC JTC 1/SC 29/WG1 FCD 14495 public draft, July 16, 1997; available at <http://www.jpeg.org/>.

Selective Image Sharpening by Simultaneous Nonlinear-Diffusion Process with Spatially Varying Parameter Presetting

Takahiro Saito, Shigemitsu Anyoji, and Takashi Komatsu

Department of Electrical, Electronics and Information Engineering,
High-Tech Research Center, Kanagawa University, Yokohama, 221-8686, Japan
{saitot01,r200370085,komatt01}@kanagawa-u.ac.jp

Abstract. Previously we have presented a PDE-based method for selective image sharpening. Our method works as the simultaneous nonlinear-diffusion process composed of a nonlinear-diffusion term, a fidelity term and a peaking term, and it sharpens blurred edges while smoothing out noisy variations. However, our method has the problem that it does not satisfactorily sharpen complex image-structures such as T-shaped edges and textures. This paper copes with the problem, and improves selective image sharpening. As the preprocess of our PDE-based method, this paper introduces a step to classify each pixel into two categories on the basis of mid-scale image-features contained in the image-gradient field. The classification results are then utilized to preset the parameters characterizing our PDE-based method spatially varyingly.

1 Introduction

For image sharpening, various filtering-based methods such as the peaking method [1] have been developed, but they have the common disadvantage that they will not work well for noisy blurred images; they will increase the noise visibility. We need to reinforce the image sharpening with the selectivity of blurred edges to be sharpened from noisy variations to be smoothed out.

For the selective image sharpening, recently, the new line has appeared; the nonlinear PDE(Partial Differential Equation)-based methods have been developed[2] -[9]. Among them, the most basic one is the bounded-variation image restoration method[3]. This method restores degraded images by minimizing the energy functional composed of the restoration energy and the smoothness energy. The time-evolving equation for the minimization is given by the Euler-Lagrange PDE, and its steady-state solution is used as a sharpened image. The method needs the accurate knowledge of the model of image blurs. If we define the restoration energy properly, the method will be able to achieve the image sharpening. However, if the model of image blurs is not accurate, it will produce some visible artifacts.

The PDE-based methods of the other category are the methods that do not need any accurate knowledge of the model of image blurs and use the transient

solution of the time-evolving PDE as a sharpened image[5]-[8]. The typical methods are the forward-and-backward diffusion method proposed by Gilboa, Sochen and Zeevi [6], and our regularized simultaneous nonlinear-diffusion method[7],[8]. In our previous paper[7],[8], we have developed the regularized simultaneous nonlinear-diffusion method, by modifying the prototypal PDE originally proposed by Proesmans et al for image restoration [9]. Our PDE-based selective image-sharpening method is a generic method that can be applied to various image-processing tasks.

Our method works as the time-evolution composed of a nonlinear-diffusion term, a fidelity term and a peaking term, and it controls the antagonism among the three componential terms so that it can sharpen only blurred edges while smoothing out noisy variations. However, our method has the problem that it does not satisfactorily sharpen complex image-structures: T-shaped edges, cross-shaped edges, textures, and so on. This paper copes with the problem. In advance of the time-evolution of the selective image sharpening, analyzing the gradient-vector field of an input noisy blurred image statistically, we classify all the pixels into the two categories: the simple-edge region and the compound region. We utilize the results of the pixel classification to preset the weighting coefficients of the three componential terms in the time-evolution, at each pixel, so that the antagonism among them can be controlled properly to each pixel in the two classification categories.

2 Regularized Simultaneous Nonlinear-Diffusion Process

As for the explicit backward-diffusion process with a negative diffusion-coefficient, the peaking method[1] is the most popular, and it can sharpen blurred edges. The peaking method creates overshoots near blurred edges by

$$f - a \cdot \Delta f \quad . \quad (1)$$

However, the backward-diffusion process has crucial drawbacks: instability, oscillations and noise amplification. The promising approach to cope with the difficulty is to introduce the antagonism between the forward diffusion and the backward diffusion into the process. We need to control the antagonism so that the backward diffusion will dominate over the forward diffusion only near blurred edges to be sharpened. The typical methods utilizing the antagonism are the forward-and-backward diffusion method proposed by Gilboa, Sochen and Zeevi[6], and our regularized simultaneous nonlinear-diffusion method[7],[8].

Our regularized simultaneous nonlinear-diffusion method introduces the antagonism between the P-M(Perona and Malik) nonlinear-diffusion process and the peaking process, and for the peaking process it uses regularized spatial-derivatives given by the simultaneous P-M process of image derivatives. The utilization of the regularized derivatives resembles the idea of the spatial regularization proposed by Catte et al[10], but instead of the Gaussian smoothing, our method adopts the P-M process to keep abrupt changes of the image derivatives. The time-evolving equations of our method are given by Equation(2), where the auxiliary functions u , v , p , q approximate the spatial derivatives of the

time-evolving image f along the horizontal (x -), vertical (y -), diagonal (d -) and anti-diagonal (a -) axes, respectively. In our prototypal scheme[7],[8], only the two auxiliary functions u, v are used, but Equation(2) utilizes the four auxiliary functions u, v, p, q to prevent the directional bias inherent in our prototypal scheme.

$$\begin{aligned}
 \frac{\partial f}{\partial \tau} &= \text{div} [c_1 (\|\nabla f\|) \cdot \nabla f] - \lambda_f \cdot (f - g) - s \cdot \left\{ \lambda_c \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) + \lambda_d \left(\frac{\partial p}{\partial d} + \frac{\partial q}{\partial a} \right) \right\}, \\
 \frac{\partial u}{\partial \tau} &= \text{div} [c_2 (\|\nabla u\|) \cdot \nabla u] - \lambda_u \cdot \left(u - \frac{\partial f}{\partial x} \right), \\
 \frac{\partial v}{\partial \tau} &= \text{div} [c_2 (\|\nabla v\|) \cdot \nabla v] - \lambda_v \cdot \left(v - \frac{\partial f}{\partial y} \right), \\
 \frac{\partial p}{\partial \tau} &= \text{div} [c_3 (\|\nabla p\|) \cdot \nabla p] - \lambda_p \cdot \left(p - \frac{\partial f}{\partial d} \right), \\
 \frac{\partial q}{\partial \tau} &= \text{div} [c_3 (\|\nabla q\|) \cdot \nabla q] - \lambda_q \cdot \left(q - \frac{\partial f}{\partial a} \right), \\
 c_1(z) &= 1 / \left(1 + (z/K)^2 \right); \quad c_2(z) = 1 / \left(1 + (\sqrt{2}z/K)^2 \right); \\
 c_3(z) &= 1 / \left(1 + (2z/K)^2 \right), \\
 \lambda_c + \lambda_d &= 1.0, \quad 0 \leq \lambda_c \leq 1.0, \\
 s &: \text{shooting parameter}, \quad \lambda_u, \lambda_v, \lambda_p, \lambda_q : \text{fidelity parameter}, \\
 f &: \text{time-evolving image}, \quad g : \text{input image}, \\
 u, v, p, q &: \text{time-evolving auxiliary functions}.
 \end{aligned}
 \tag{2}$$

The first equation in Equation(2) formulates the time-evolution of the image f , and is composed of the P-M nonlinear-diffusion term, the fidelity term and the peaking term. The amount of overshoots to be added is controlled by the shooting parameter s that is set to a positive value. In smoothly-varying image regions, the P-M term defeats the peaking term, which results in smoothing out noisy variations; whereas near blurred edges the peaking term defeats the P-M term, which results in producing overshoots. All the equations except the first equation formulate the time-evolution for regularizing the four auxiliary functions u, v, p, q and they are composed of the P-M term and the fidelity term. Their time-evolving auxiliary functions u, v, p, q are used for the computation of the peaking term. The discrete algorithm for the first equation is given by Equation(3).

$$\begin{aligned}
 f_{i,j}^{\tau+1} &= f_{i,j}^{\tau} + \varepsilon \cdot \left[\lambda_c \sum_{d=N,S,E,W} [c_1 (\|\nabla_d f_{i,j}^{\tau}\|) \cdot \nabla_d f_{i,j}^{\tau}] \right. \\
 &+ \frac{\lambda_d}{2} \sum_{d=SE,NW,SW,NE} \left[c_1 \left(\frac{1}{\sqrt{2}} |\nabla_d f_{i,j}^{\tau}| \right) \cdot \nabla_d f_{i,j}^{\tau} \right] - \lambda_f \cdot (f_{i,j}^{\tau} - g_{i,j}) \\
 &\left. - s \cdot \left\{ \frac{\lambda_c}{2} \cdot [(u_{i+1,j}^{\tau+1} - u_{i-1,j}^{\tau+1}) + (v_{i,j+1}^{\tau+1} - v_{i,j-1}^{\tau+1})] \right\} \right] \\
 &+ \frac{\lambda_d}{2\sqrt{2}} \cdot [(p_{i+1,j+1}^{\tau+1} - p_{i-1,j-1}^{\tau+1}) + (q_{i-1,j+1}^{\tau+1} - q_{i+1,j-1}^{\tau+1})] \Big\}, \\
 \nabla_N f_{i,j}^{\tau} &= f_{i,j-1}^{\tau} - f_{i,j}^{\tau}, \quad \nabla_S f_{i,j}^{\tau} = f_{i,j+1}^{\tau} - f_{i,j}^{\tau}, \\
 \nabla_E f_{i,j}^{\tau} &= f_{i+1,j}^{\tau} - f_{i,j}^{\tau}, \quad \nabla_W f_{i,j}^{\tau} = f_{i-1,j}^{\tau} - f_{i,j}^{\tau}, \\
 \nabla_{SE} f_{i,j}^{\tau} &= f_{i+1,j+1}^{\tau} - f_{i,j}^{\tau}, \quad \nabla_{NW} f_{i,j}^{\tau} = f_{i-1,j-1}^{\tau} - f_{i,j}^{\tau}, \\
 \nabla_{SW} f_{i,j}^{\tau} &= f_{i-1,j+1}^{\tau} - f_{i,j}^{\tau}, \quad \nabla_{NE} f_{i,j}^{\tau} = f_{i+1,j-1}^{\tau} - f_{i,j}^{\tau}, \\
 f_{i,j}^{\tau}, u_{i,j}^{\tau}, v_{i,j}^{\tau}, p_{i,j}^{\tau}, q_{i,j}^{\tau} &: \text{values of } f, u, v, p, q \text{ on the pixel location } (i, j) \\
 &\text{at the } \tau\text{-th iteration}.
 \end{aligned}
 \tag{3}$$

As a decision scheme to halt its iteration, we employ the following scheme:

$$\begin{aligned}
 & \text{If } |\delta^{\tau-1, \tau} - \delta^{\tau, \tau+1}| \leq \delta_T, \text{ then stop the iteration.} \\
 & \delta^{\tau, \tau+1} = \delta_f^{\tau, \tau+1} / \delta_w^{\tau+1}, \\
 & \delta_f^{\tau, \tau+1} = \sum_{i,j} |f_{i,j}^{\tau+1} - f_{i,j}^{\tau}|, \delta_w^{\tau+1} = \sum_{i,j} (|u_{i,j}^{\tau+1}| + |v_{i,j}^{\tau+1}| + |p_{i,j}^{\tau+1}| + |q_{i,j}^{\tau+1}|) / 4.
 \end{aligned} \tag{4}$$

The decision scheme will halt its iteration almost at the ideal moment when it achieves the best selective image sharpening.

3 Parameter Presetting Based on Pixel Classification

If our method can adjust the antagonism among its three componential terms to each pixel, its capability of selective sharpening will be improved. For the automatic adjustment, in advance of the time-evolution, the parameters included in the underlying PDE are preset spatially varyingly to their proper values that are regarded as depending on mid-scale image-features. The mid-scale features correspond to some features between global image-features and local ones, and they characterize coherent image regions. This paper defines the intermediate features as the mid-scale statistics of the vector field of image gradients, and on the basis of the mid-scale gradient-statistics we define the two categories of image regions: the simple-edge region and the compound region. In advance of the time-evolution, analyzing the gradient-vector field of a given noisy blurred image statistically, we classify all the pixels into the two categories, and then preset the parameters included in the underlying PDE at each pixel in the two categories so that the antagonism among its three componential terms will be controlled for the selective sharpening.

3.1 Pixel Classification

First, at each pixel, we compute the gradient of a given noisy blurred image $g(x, y)$. The local image-gradients are calculated by a correlation of masks with the image data. This paper employs the masks for the moment-based edge detector. The gradient magnitude R and direction φ are defined by

$$R = |\nabla g| \quad ; \quad \varphi = \arctan \left(\frac{\partial g}{\partial y} / \frac{\partial g}{\partial x} \right) \tag{5}$$

Next, a statistical procedure is applied to the estimated local image-gradients, to obtain the mid-scale statistics of the gradient field. Assuming that the pixel value $g_{i,j}$ of each pixel (i, j) is modeled by the sum of a image-structure part and a random white Gaussian noise of zero mean and standard deviation σ_e , the probability density function $P_{i,j}(\theta)$ of the gradient direction θ within the range $(-\pi/2, \pi/2]$ is given by

$$\begin{aligned}
 p_{i,j}(\theta) = \exp \left(-\frac{(r_{i,j})^2}{2} \right) \cdot \left[\frac{1}{\pi} + \frac{2r_{i,j} \cdot \cos(\theta - \varphi_{i,j})}{\sqrt{2\pi}} \times \right. \\
 \left. \exp \left(\frac{(r_{i,j})^2 \cdot \cos^2(\theta - \varphi_{i,j})}{2} \right) \cdot \text{erf} \left(r_{i,j} \cdot \cos(\theta - \varphi_{i,j}) \right) \right] \quad ; \quad r_{i,j} = \frac{R_{i,j}}{\sigma} \tag{6}
 \end{aligned}$$

where σ is the standard deviation of noise mixed in the x and y components of the calculated gradient, and $r_{i,j}$ is the normalized gradient magnitude [11]. We define the gradient-direction profit by summing up the probability density function within the neighboring region centered at pixel (i, j) ,

$$\bar{p}_{i,j}(\theta) = \frac{1}{25} \cdot \sum_{k=-2}^2 \sum_{l=-2}^2 p_{i+k,j+l}(\theta) \quad (7)$$

The profit provides information of directional tendencies within the neighboring region. The maximization of the profit with respect to the gradient direction θ gives an improved estimate of the gradient direction at the pixel (i, j) . In practice, the profit is computed for the eight angles:

$$\theta = \left\{ -\frac{3\pi}{8}, -\frac{\pi}{4}, -\frac{\pi}{8}, 0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2} \right\} \quad (8)$$

Among them, we select the best angle $\theta_{O(i,j)}$ giving the maximum profit, as the dominant gradient direction. However, if the value of the maximum profit is small and is under a threshold $T_P (= 0.5)$, then the selected angle is not considered reliable and the dominant gradient-direction $\theta_{O(i,j)}$ is set to 'nil'. The 'nil' value means that there exists no dominant gradient-direction within the neighboring region.

In addition to $\theta_{O(i,j)}$, the average normalized gradient-magnitude $\mu_{i,j}$ is computed within the region centered at an pixel (i, j) ,

$$\mu_{i,j} = \frac{1}{81} \cdot \sum_{k=-4}^4 \sum_{l=-4}^4 r_{i+k,j+l} \quad (9)$$

and then $\mu_{i,j}$ is transformed to the measure of intensity variation $\rho_{i,j}$,

$$\rho_{i,j} = 1 - 1 / \left\{ 1 + (\mu_{i,j} / 1.5)^2 \right\} \quad (10)$$

The value of $\rho_{i,j}$ is in the range $[0, 1]$. As $\mu_{i,j}$ increases, $\rho_{i,j}$ approaches one. Utilizing the results of the above statistical analysis of the gradient field of an input image, we classify all the pixels into the two categories: the simple-edge region and the compound region, as follows:

(1) Simple-edge region: Definition: $\{ \theta_{O(i,j)} \neq nil \}$ (11)

(2) Compound region: Definition: $\{ \theta_{O(i,j)} = nil \}$ (12)

Figure1 shows an input noisy blurred test image and its classification result. The pixel classification provides the reasonable result even for noisy blurred images.

3.2 Spatially Varying Parameter Presetting

For each pixel (i, j) , we preset the parameters $s_{i,j}$, $\lambda_{f(i,j)}$, $\lambda_{c(i,j)}$, $\lambda_{d(i,j)}$ to their proper values.

(1) Simple-edge region:

Preset the parameters so that the edge sharpening will have priority over the noise smoothing. The shooting parameter $s_{i,j}$ and the fidelity parameter $\lambda_{f(i,j)}$ are preset to their proper constant values and the parameters $\lambda_{c(i,j)}$, $\lambda_{d(i,j)}$ are



(a) Noisy blurred image (PSNR=25.346dB) (b) Classification result White pixels: Simple-edge region

Fig. 1. Pixel classification

defined as the functions of the estimated dominant gradient-direction $\theta_{O(i,j)}$, as follows:

$$s_{i,j} = 0.6; \lambda_{f(i,j)} = 0.5; \lambda_{c(i,j)} = \begin{cases} 1, & \text{if } \theta_{O(i,j)} = 0, \frac{\pi}{2} \\ 0, & \text{if } \theta_{O(i,j)} = -\frac{\pi}{4}, \frac{\pi}{4} \\ \frac{1}{2}, & \text{otherwise} \end{cases} \quad (13)$$

$$\lambda_{d(i,j)} = 1 - \lambda_{c(i,j)}.$$

(2) *Compound region:*

Preset the parameters so that complex image-structures will be sharpened in complex-structure image portions whereas the noise smoothing will have priority over the edge sharpening in smooth-gradation image portions. The shooting parameter $s_{i,j}$ and the fidelity parameter $\lambda_{f(i,j)}$ are defined as the functions of the estimated intensity variation $\rho_{i,j}$, as follows:

$$s_{i,j} = 0.6 \cdot \rho_{i,j}; \lambda_{f(i,j)} = 0.3 \cdot \rho_{i,j} + 0.5; \lambda_{c(i,j)} = 1.0; \lambda_{d(i,j)} = 0.0. \quad (14)$$

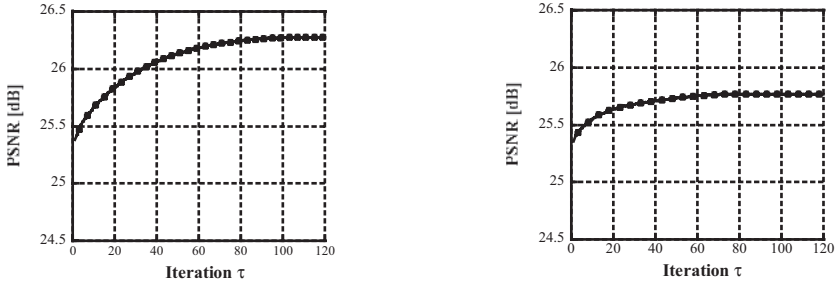
The parameter presetting is heuristic in nature, and hence we have determined the proper values through experiments using many noisy blurred test images.

4 Performance Evaluations

We evaluate performance of our method using artificially blurred test images. First we blur an original sharp image $h(x, y)$ with the Gaussian filter having the impulse response $G(x, y; \zeta)$,

$$G(x, y; \zeta) = \frac{1}{2\pi\zeta^2} \cdot \exp\left(-\frac{(x^2+y^2)}{2\zeta^2}\right), \quad (15)$$

and then add random Gaussian noise $n(x, y)$ with zero mean and standard deviation of 5.0. We set the blurring parameter ζ of the Gaussian filter to 1.0. For the performance evaluation, we compute PSNR[dB] between the original sharp image h and the sharpened image f .



(a) Spatially varying parameter presetting (b) Spatially uniform parameter presetting

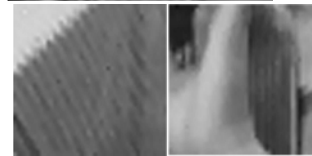
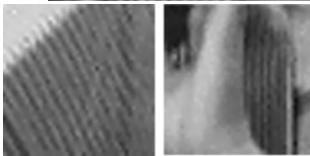
Fig. 2. Update characteristics of our PDE-based method



(a) Part of the blurred test image (PSNR=24.346[dB])



(b) Part of the sharped image by our method using the spatially varying parameter presetting (PSNR=26.273[dB])



(c) Spatially varying parameter presetting (d) Spatially uniform parameter presetting

Fig. 3. Enlarged parts of the sharped images by our PDE-based method

As for the parameter K , its proper value chiefly depends on the noise variance. If we set the parameter K close to the standard deviation of noise, our scheme almost provides optimal performance. We set the parameters of ε, δ_T and K to 0.05, 10^{-4} and 5.0, respectively; and these optimal values are determined by experiments. In addition, we fix all the parameters $\lambda_u, \lambda_v, \lambda_p, \lambda_q$ at equally 1.0.

We apply our spatially varying parameter presetting to the selective sharpness-enhancement of the noisy blurred test image shown in Fig.1(a). Figure 2 compares the update characteristics of our PDE-based sharpening method using the spatially varying parameter presetting to those using the existing spatially uniform parameter presetting where the parameters $s, \lambda_f, \lambda_c, \lambda_d$ are set to their optimal values. In this case, the application of the decision scheme halts the time-evolution of our sharpening method using the spatially varying parameter presetting at the 100-th iteration where the PSNR is almost the highest. Figure 3 compares enlarged parts of the sharpened image by our PDE-based methods. In the case of the spatially uniform presetting some complex image structures disappear, whereas in the case of the spatially varying presetting they are preserved and enhanced to some extent.

References

1. A. Rosenfeld and A.C. Kak: Digital picture processing, Ch.6.4.2., Academic Press, Inc., New York, 1982
2. G. Aubert and P. Kornprobst: Mathematical problems in image processing - Partial differential equations and the calculus of variations, Springer Verlag, New York, 2002.
3. G. Auber and L. Vese: A variational method in image recovery, *SIAM J. Num. Anal.*, **34**, 5, pp.1948-1979, Oct. 1997.
4. P. Perona and J. Malik: Scale-space and edge detection using anisotropic diffusion, *IEEE Trans. Pattern. Anal. & Mach. Intell.*, **12**, 7, pp.629-639, July 1990.
5. S. Osher and L.I. Rudin: Feature-oriented image enhancement using shock filters, *SIAM J. Num. Anal.*, **27**, 4, pp.919-940, Aug 1990.
6. G. Gilboa, N. Sochen and Y. Zeevi,: Forward-and-backward diffusion processes for adaptive image enhancement and denoising, *IEEE Trans. Pattern. Image Process.*, **11**, 7, pp.689-703, July 2002.
7. T. Saito, J. Satsumabayashi, K. Yashiro and T. Komatsu,: Selective image sharpness enhancement by coupled nonlinear reaction-diffusion time-evolution and its practical application, *Proc. EUSIPCO 2002*, vol.II, pp.445-448, Toulouse, France, Sept. 2002.
8. T. Saito, H. Harada, J. Satsumabayashi and T. Komatsu,: Color image sharpening based on nonlinear reaction-diffusion, *Proc. IEEE ICIP*, vol.III, pp.389-392, Barcelona, Spain, Sept. 2003.
9. M. Proesmans, E.J. Pauwels and L.J. Van Gool: Coupled geometry driven diffusion equations for low-level vision, *Geometry-Driven Diffusion in Computer Vision*, B.M. ter Haar Romeny (Ed.), pp.191-228, Kluwer, Dordrecht, 1994.
10. F. Catte, P. L. Lions, J.M. Morel and T. Coll: Image selective smoothing and edge detection by nonlinear diffusion, *SIAM J. Num. Anal.*, **29**, 1, pp.182-193, Feb. 1992.
11. E.P. Lyvers and O.R. Mitchell,: Precision edge contrast and orientation estimation, *IEEE Trans. Pattern. Anal. & Mach. Intell.*, **10**, 6, pp.927-937, Nov. 1988.

Directional Weighting-Based Demosaicking Algorithm*

Tsung-Nan Lin^{1,2} and Chih-Lung Hsu²

¹ Dept. of Electrical Engineering,

² Graduate Institute of Communication Engineering,
National Taiwan University, Taipei, Taiwan
tsungnan@ntu.edu.tw,

Abstract. This paper presents a new color interpolation method for Color Filter Array that is commonly used in a single-sensor digital camera. We propose a spatial directional interpolation scheme that makes use of the local geometric information extracted from the surrounding pixels. Based on the image model that R, G, B channels are highly correlated, the weighting-based interpolation is performed in the color difference domain instead of in the original pixel domain. The weights, determined based on the approach of directional-based signal distance, are capable of adapting to the local image structure. Experimental results indicate the proposed method can reconstruct the missing pixel values better and preserve the sharp edge information at the same time without color-bleeding artifacts. Both subjective visual evaluation and objective performance measurement show the proposed method outperforms many existing methods.

1 Introduction

Commercially available Digital Still Cameras (DSCs) have been widely used as input devices for multimedia systems and expected to replace traditional cameras. Digital Still Cameras capture color information by using three color filters of an image sensor (CCD or CMOS). The sensor surface is covered with a mosaic of color filters, each point capturing only one sample of the color spectrum. This kind of sensor is called Color Filter Array (CFA).

Fig. 1 (a) shows one popular CFA pattern used today. It is called Bayer pattern [4]. In this arrangement, each pixel on the sensor samples only one of the three color values. Green pixels (the luminance channel) are sampled at higher rate than red and blue pixels (the chrominance channels). To obtain full-resolution color images, missing color values at each pixel must be estimated from neighboring samples. This process is known as *demosaicking*. If demosaicking is not performed appropriately, images suffer from highly visible color artifacts.

Many methods have been proposed to estimate the unknown color values. The most basic idea is to independently interpolate the R, G, and B planes using linear interpolation. Despite its simplicity, such linear interpolation scheme

* This work is supported by NSC, Taiwan grant NSC92-2622-E-002-011-CC3

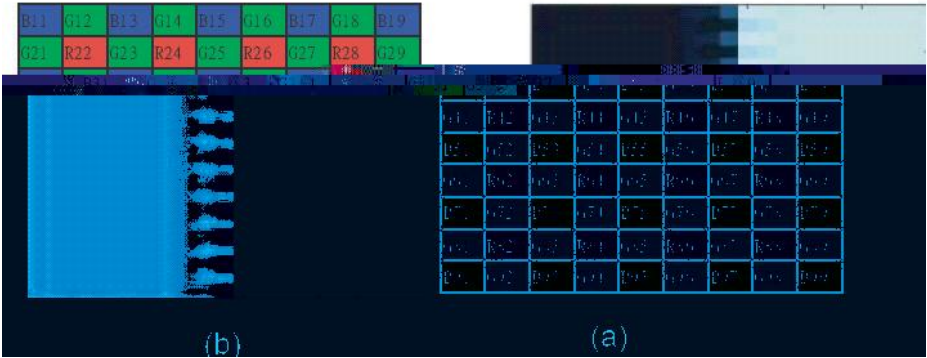


Fig. 1. (a) A CFA image. (b) The zipper artifact.

often introduces visible color bleeding or color aliasing artifacts around sharp or fine edges. This is referred to as the zipper effect [1,2] as shown in Figure 1 (b). The key objection to pixel artifacts in images that result from linear interpolation is abrupt and unnatural hue change. Cox [7,8] observed a simple spectral correlation between different color planes. Algorithms [7,8,12] hypothesize that the quotient of two color channels ($\frac{R}{G}, \frac{B}{G}$) is slowly varying within a local image region. By interpolating the hue value, hues are allowed to change only gradually. An alternate approach is to interpolate the difference between red and green and blue and green [3,9,6]. Moreover, schemes [5,13,14,10] first detect the spatial content in a local region and are able to interpolate the missing color values more effectively. However, these schemes often fail to perform well on sharp edges.

This paper presents an improved demosaicing algorithm based on a directional weighting-based interpolation scheme. According to the local image structure, the weights are determined based on the approach of directional-based signal distance. The idea is to interpolate along the edge boundary instead of across the boundary.

2 Weighted-Sum Approach

Natural images possess strong spatial correlations within a local region where neighboring pixels share similar color values. Many sophisticated demosaicing algorithms have been proposed to exploit image spatial correlation and interpolate the missing color pixels by the weighted-sum approach. Representative method includes Kimmel’s method [12] which proposes to use the gradient information as an edge indicator. The interpolation is achieved by weighting neighboring pixels. However, existing methods which do not exploit spatial information well often result in zipper-effect artifact around edges. The main drawbacks of existing approaches are they can only detect vertical and horizontal edges efficiently but unable to handle edges with 45 or 135 degrees. We presents the following two simple image analysis based on Kimmel’s method to demonstrate

the problem. The interpolation of a missing G_{ij} pixel can be accomplished by weighting neighboring pixels as

$$G_{ij} = \frac{w_{i-1j}G_{i-1j} + w_{i+1j}G_{i+1j} + w_{ij-1}G_{ij-1} + w_{ij+1}G_{ij+1}}{w_{i-1j} + w_{i+1j} + w_{ij-1} + w_{ij+1}} \tag{1}$$

Kimmel defined the weighting factor w as

$$w_{i+1j} = (1 + Dx_{ij}^2 + Dx_{i+1j}^2)^{1/2} \tag{2}$$

$$Dx_{ij} = \frac{P_{i+1j} - P_{i-1j}}{2}, Dy_{ij} = \frac{P_{ij+1} - P_{ij-1}}{2} \tag{3}$$

where P_{ij} denotes the value of single color pixel at location (i, j) , and Dx_{ij} and Dy_{ij} represent the gradient approximation in horizontal and vertical directions respectively.

Let's see the following two demosaicking examples of a vertical edge (Figure 2(a)) and an edge of 45 degree (Figure 2(b)). To calculate the weighted interpolation of $G(7)$ in Figure 2(a), the set of weights is calculated as:

$$w1 = \frac{1}{(1 + (G11 - G13)^2 + (B7 - B2)^2)^{1/2}} = \frac{1}{(1 + (0 - 0)^2 + (0 - 0)^2)^{1/2}} = 1$$

$$w2 = \frac{1}{(1 + (G8 - G6)^2 + (B7 - B5)^2)^{1/2}} = \frac{1}{(1 + (200 - 0)^2 + (0 - 0)^2)^{1/2}} = 0.005$$

$$w3 = \frac{1}{(1 + (G6 - G8)^2 + (B7 - B9)^2)^{1/2}} = \frac{1}{(1 + (0 - 200)^2 + (0 - 200)^2)^{1/2}} = 0.0035$$

$$w4 = \frac{1}{(1 + (G3 - G11)^2 + (B7 - B12)^2)^{1/2}} = \frac{1}{(1 + (0 - 0)^2 + (0 - 0)^2)^{1/2}} = 1$$

The interpolated value of $G7$ is

$$G7 = \frac{w1 * G3 + w2 * G6 + w3 * G8 + w4 * G11}{w1 + w2 + w3 + w4} = 0.3485$$

In the case of Figure 2 (b), the weights become $w1 = 0.005, w2 = 0.005, w3 = 0.0035, w4 = 0.0035$. The interpolated $G7$ becomes

$$G7 = \frac{w1 * G3 + w2 * G6 + w3 * G8 + w4 * G11}{w1 + w2 + w3 + w4} = 117$$

It can be easily seen the tradition weighted sum approach is capable of exploiting either horizontal or vertical edge efficiently, but fails to detect edges with 45-degree (or 135-degree).

3 An Adaptive Spatial Filter of Detecting All Orientation Edges

In this section, we propose an adaptive filtering scheme which is capable of detecting edges with any orientations. Our approaches are based on a non-cross

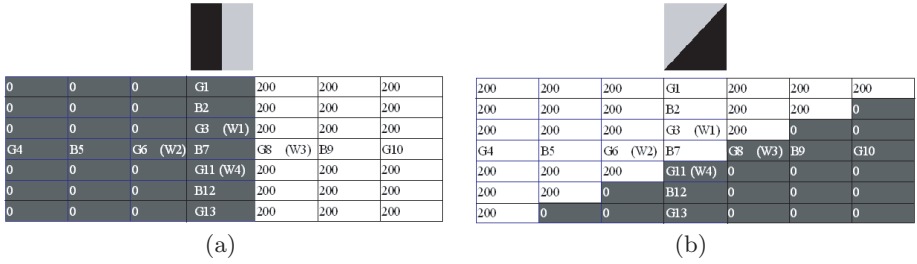


Fig. 2. (a) A CFA image with vertical edge. (b) A CFA image with 45-degree edge.

weighting mechanism which estimates the similarity correlation of the neighboring pixels to the central pixel. Our idea of determining directional weights is inspired by various edge-directed demosaicking schemes [2,7,11,14,10], which perform color interpolation along rather than across the image edges by first analyzing the spatial structure of a local image neighborhood. The north weight $w1$ in Figure 2 is determined inversely proportional to the signal difference along this direction to the center pixel. We hypothesis that the weighting function $f(\cdot)$ of detecting the local structure is of $G(3)$ inversely proportional to the signal difference $\delta(G3, G7)$. Since $G(7)$ is unknown, we can approximate $f(\delta(G3, G7)) \approx f(\delta(G1, G3)) + f(\delta(B2, B7)) \approx f(\delta(G1, G3) + \delta(B2, B7))$ if we assume the single difference in local structure to be kept constant. Therefore, the weighting function of $w1$ can be designed as $\frac{1}{(1+(\delta(G1, G3) + \delta(B2, B7))^2)^{1/2}}$ in a 5×5 window. In a 7×7 window, the weights can be calculated as

$$w1 = \frac{1}{(1 + (G1 - G3)^2 + (B7 - B2)^2)^{1/2}} \tag{4}$$

$$w2 = \frac{1}{(1 + (G4 - G6)^2 + (B7 - B5)^2)^{1/2}} \tag{5}$$

$$w3 = \frac{1}{(1 + (G10 - G8)^2 + (B7 - B9)^2)^{1/2}} \tag{6}$$

$$w4 = \frac{1}{(1 + (G13 - G11)^2 + (B7 - B12)^2)^{1/2}} \tag{7}$$

When applying the proposed non-cross weighting function to the CFA image with the vertical edge in Figure 2(a), we can get $w1 = 1, w2 = 1, w3 = 0.005, w4 = 1$ and the interpolated $G(7) = 0.1664$. In the case of Figure 2(b), the corresponding weights are calculated as $w1 = 1, w2 = 1, w3 = 0.005, w4 = 0.005$ and the interpolated $G(7) = 199$. Figure 3 (a) shows the error analysis of traditional weight function [12] in detecting edges with 45-degree (or 135 degree). The y axis is the interpolated error percentage to the signal difference (x axis). It can be easily seen that the interpolated error results in the zipper artifacts in the demosaicked image. On the other hand, the proposed approach is capable of detecting local structure with all orientations as shown in Figure 3 (b). Based on the proposed adaptive spatial filter, the demosaicking algorithm is described in details in Section 4.

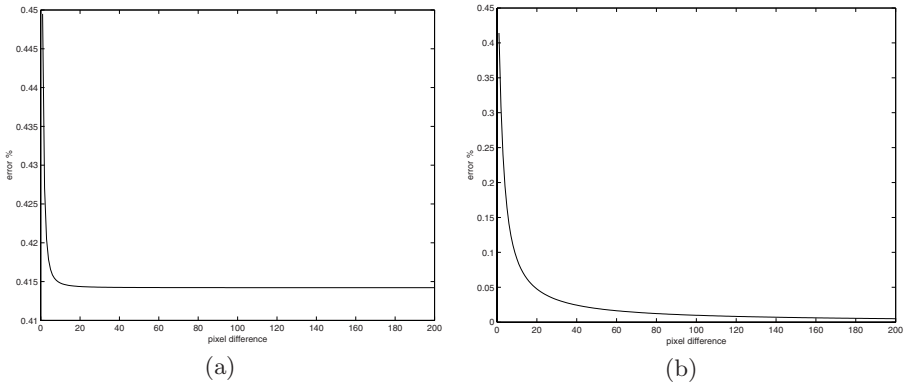


Fig. 3. Error analysis of edge with 45 degree. *y* axis is the error percentage to the signal difference between the edge. *x* axis is the signal difference between the edge. (a) Results of Kimmel. (b) Results of the proposed filter.

4 Directional Weighting-Based Interpolation Method

The image model proposed by Adams [2,3] assumes red and blue channels are perfectly correlated to green channels over the extent of the interpolation pixel neighborhood. The image model can be described as

$$G(n) = R(n) + k_R(n) \tag{8}$$

$$G(n) = B(n) + k_B(n) \tag{9}$$

where *n* refers to the pixel location. *k_B* and *k_R* is defined as the difference between green and blue, and green and red respectively. For real-world images, the contrast of *k_B* and *k_R* is quite flat over a local image region. Based on this observation, it is reasonable to estimate the missing pixel values in the *k_B* and *k_R* domain instead of pixel value domain.

The proposed demosaicking algorithm first estimates the missing G values at R and B pixels. For instance to estimate the missing green value of $\hat{G}(44)$ shown in Figure 1, we need to reconstruct the corresponding $\hat{k}_R(44)$. Then the missing green value of $\hat{G}(44)$ can be obtained:

$$\hat{G}(44) = R(44) + \hat{k}_R(44) \tag{10}$$

where $\hat{k}_R(44)$ can be estimated based on weighting-based interpolation of the set $\{k_R(34), k_R(43), k_R(54), k_R(45)\}$.

$$\hat{k}_R(44) = \frac{\sum_{i=34,43,45,54} w_{k_R(i)} * k_R(i)}{\sum_{i=34,43,45,54} w_{k_R(i)}} \tag{11}$$

Weight $w_{k_R}(34), w_{k_R}(43), w_{k_R}(45)$, and $w_{k_R}(54)$ represent the similarity measurements of the corresponding values of *k_R*(*i*) to the central pixel *k_R*(44). We

use the directional-based single distance approach to estimate those weights based on the following equations:

$$w_{k_R}(34) = \frac{1}{\sqrt{1 + (R(44) - R(24))^2 + (G(34) - G(14))^2}} \tag{12}$$

$$w_{k_R}(43) = \frac{1}{\sqrt{1 + (R(44) - R(42))^2 + (G(43) - G(41))^2}} \tag{13}$$

$$w_{k_R}(54) = \frac{1}{\sqrt{1 + (R(44) - R(64))^2 + (G(54) - G(74))^2}} \tag{14}$$

$$w_{k_R}(45) = \frac{1}{\sqrt{1 + (R(44) - R(46))^2 + (G(45) - G(47))^2}} \tag{15}$$

Since no R values are available at the locations of 34, 43, 45, 54, a simple linear prediction scheme is used to estimate $\hat{k}_R(\cdot)$ values.

$$\hat{k}_R(34) = G(34) - \frac{R(44) + R(24)}{2} \tag{16}$$

$$\hat{k}_R(43) = G(43) - \frac{R(44) + R(42)}{2} \tag{17}$$

$$\hat{k}_R(54) = G(54) - \frac{R(44) + R(64)}{2} \tag{18}$$

$$\hat{k}_R(45) = G(45) - \frac{R(44) + R(46)}{2} \tag{19}$$

The steps to estimate the missing G values at B pixels are similar.

Then we estimate the missing B values at R pixels and the missing R values at B pixels. For example, to estimate $\hat{B}(44)$ at the location 44 (where $R(44)$ is the sensor capture value and $\hat{G}(44)$ is estimated already), $\hat{k}_B(44)$ is calculated based on the weighted-interpolation

$$\hat{k}_B(44) = \frac{\sum_{i=33,53,55,35} w_{k_B}(i) * k_B(i)}{\sum_{i=33,53,55,35} w_{k_B}(i)}. \tag{20}$$

Then the value of $B(44)$ can be estimated based on the equation:

$$\hat{B}(44) = \hat{G}(44) - \hat{k}_B(44) \tag{21}$$

The set of weights can be calculated based on the directional signal strength similarity to the central pixel of $R(44)$ accordingly.

$$w_{k_B}(33) = \frac{1}{\sqrt{1 + (R(44) - R(22))^2 + (B(33) - B(11))^2}} \tag{22}$$

$$w_{k_B}(53) = \frac{1}{\sqrt{1 + (R(44) - R(62))^2 + (B(53) - B(71))^2}} \tag{23}$$

$$w_{k_B}(55) = \frac{1}{\sqrt{1 + (R(44) - R(66))^2 + (B(55) - B(77))^2}} \tag{24}$$

$$w_{k_B}(35) = \frac{1}{\sqrt{1 + (R(44) - R(26))^2 + (B(35) - B(17))^2}} \tag{25}$$

Table 1. The objective PSNR performance (in dB) of different demosaic algorithms.

	Bilinear	[7]	[1]	[14]	[10]	[5]	Proposed
Window	34.75	36.58	36.73	38.93	40.48	40.58	42.65
Sailboat	33.88	36.04	36.86	38.56	39.79	39.21	42.72
Statue	33.28	35.31	34.93	36.4	38.16	39.2	41.59
Lighthouse	29.56	31.71	33.42	35.47	36.1	33.82	40.31

The procedures to estimated the missing B values at R pixels can be applied similarly.

The final stage is to estimate the missing R and B values at G pixels. For instance, to reconstruct $\hat{R}(45)$, the corresponding $\hat{k}_R(45)$ can be interpolated:

$$\hat{k}_R(45) = \frac{\sum_{i=44,55,46,35} w_{k_R}(i) * k_R(i)}{\sum_{i=44,55,46,35} w_{k_R}(i)}. \quad (26)$$

where $k_R(44) = \hat{G}(44) - R(44)$, $k_R(55) = \hat{G}(55) - \hat{R}(55)$, $k_R(46) = \hat{G}(46) - R(46)$, and $k_R(35) = \hat{G}(35) - \hat{R}(35)$. After $\hat{k}_R(45)$ is calculated, $\hat{R}(45)$ can be easily recovered based on Equation 27.

$$\hat{R}(45) = G(45) - \hat{k}_R(45) \quad (27)$$

The weights are calculated as follows:

$$w_{k_R}(44) = \frac{1}{\sqrt{1 + (G(45) - G(43))^2 + (R(44) - R(42))^2}} \quad (28)$$

$$w_{k_R}(55) = \frac{1}{\sqrt{1 + (G(45) - G(65))^2 + (B(55) - B(75))^2}} \quad (29)$$

$$w_{k_R}(46) = \frac{1}{\sqrt{1 + (G(45) - G(47))^2 + (R(46) - R(48))^2}} \quad (30)$$

$$w_{k_R}(35) = \frac{1}{\sqrt{1 + (G(45) - G(25))^2 + (B(35) - B(15))^2}} \quad (31)$$

The estimated R and B values can be re-applied to Equation 16~Equation 19 to get the more accurate values of $\hat{k}_R(\cdot)$. Then Equation 11 and Equation 10 are applied again to reconstruct G values at R and B pixels.

5 Experiment Results

Four benchmark images from the Kodak photo sampler as shown in Figure 4 are chosen to evaluate the performance of the proposed algorithm. These images are first sampled with Bayer CFA to produce mosaic images, and then used as the input for interpolation. The PSNR (peak signal-to-noise ratio) measure is used to evaluate the objective performance of the proposed algorithm as compared to other existing methods [1,7,14,10,5]. The PSNR results averaged over three



Fig. 4. Four benchmark images in Kodak photo sampler. (a) Window (b) Sailboat (c) Statue (d) Lighthouse.

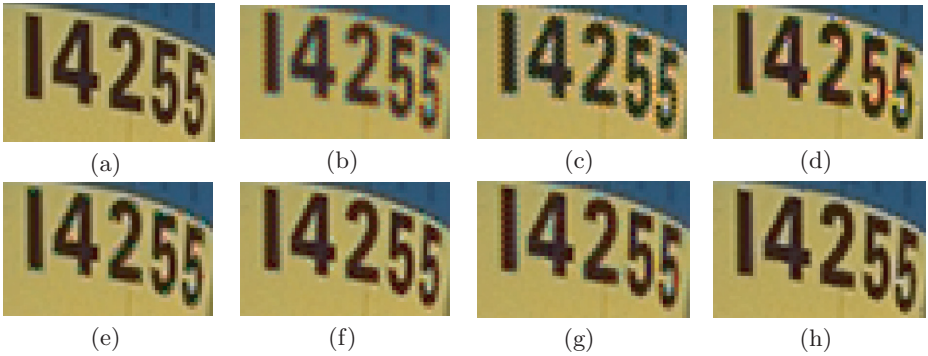


Fig. 5. Magnified portion of sailboat image. (a) The original image. (b) Method of bilinear interpolation. (c) Method of [7]. (d) Method of [1]. (e) Method of [14]. (f) Method of [10]. (g) Method of [5]. (h) The proposed Method.

color planes are reported in Table 1. The objective performance results show the proposed algorithm is significantly better than the conventional methods. In order to compare the visual quality, we magnify the interpolation results to show the details. As shown in Figure 5, the proposed algorithm generates sharp color edges without color bleeding artifacts. We can see other conventional methods produce strong zipper artifacts in the regions of fine color edges. Either the subjective visual evaluation or the objective performance measure demonstrates the superiority of the proposed method.

6 Conclusion

In this paper we present a new method CFA interpolation. This method is based on the directional weighted-based interpolation scheme. Based on the image

model that R, G, B channels are highly correlated, the interpolation is performed in the color difference domain instead of in the original pixel domain. The weights, determined based on the approach of directional-based signal distance, are capable of adapting to the local image structure. Experimental results indicate the proposed method can reconstruct the missing pixel values better and preserve the sharp edge information without color-bleeding artifacts. Both subjective visual evaluation and objective performance measurement show the proposed method outperforms many existing methods.

References

1. J. E. Adams. Interactions between color plane interpolation and other image processing functions in electronic photography. In *Proceeding SPIE 2416*, pages 144–151, 1995.
2. J. E. Adams. Design of practical color filter array interpolation algorithm for digital cameras. In *Proceeding of SPIE*, volume 3028, pages 117–125, 1997.
3. J. E. Adams. Design of color filter array interpolation algorithm for digital cameras, part 2. In *IEEE Proc. Int. Conf. Image Processing*, volume 1, pages 488–492, 1998.
4. B.E. Bayer. Color imaging array. *U.S. Patent*, No. 3,971,065, 1976.
5. E. Chang. Color filter array recovery using a threshold-based variable number of gradients. In *Proceedings of SPIE, 3650*, pages 36–43, 1999.
6. Soo chang Pei and Lo kuong Tam. Effective color interpolation in ccd color filter array using signal correlation. *IEEE ICIP*, 41, 2000.
7. D.R. Cox. Singnal processing method and apparatus for producing interpolated chrominance values in a sampled color image signal. *U.S. Patent*, No. 4,642,678, 1987.
8. D.R. Cox. Reconstruction of ccd images using template matching. In *Imaging Science & Technology, 47th Annual Conference*, pages 380–385, 1994.
9. W.T. Freeman. Median filter for reconstructing missing color samples. *U.S. Patent*, No. 4,724,395, 1988.
10. J. Hamilton and J. Adams. Adaptive color plane interpolation in single sensor color electronic camera. *U.S. Patent*, No. 5,629,734, 1997.
11. R.H. Hibbard. Apparatus and method for adaptively interpolation a full color image utilizing luminance gradients. *U.S. Patent*, No. 5,382,976, January, 1995.
12. R. Kimmel. Demosaicking: image reconstruction from color ccd sample. *IEEE trans. Image Process*, 7(3):1221–1228, 1999.
13. T. Kuno, H. Sugiura, and N. Matoba. New interpolation method using discriminated color correlation for digital still cameras. *IEEE Trans. Consumer Electronics*, 45(1):259–267, 1999.
14. C.A. Laroche and M. Prescott. Apparatus and method for adaptively interpolating in a full color image utilizing chrominance gradients. *U.S. Patent*, No. 5,373,322, 1994.

A New Text Detection Algorithm in Images/Video Frames

Qixiang Ye and Qingming Huang

Digital Media Lab, Institute of Computing Technology, and Research Center of
Digital Media, Graduate School of of Sciences, Beijing 100039, China
{qxye, qmhuang}@jd1.ac.cn

Abstract. In this paper, we propose a new text detection algorithm for images/video frames in a coarse-to-fine framework. Firstly, in the coarse detection, multiscale wavelet energy feature is employed to locate all possible text pixels and then a density-based region growing method is developed to connect these pixels into text lines. Secondly, in the fine detection, four kinds of texture features are combined to represent a text line and a SVM classifier is employed to identify texts from the candidate ones. Experimental results on two datasets show the encouraging performance of the proposed algorithm.

1 Introduction

Text contains semantic information and can do significant contribution to video database retrieval and image understanding tasks. For example, texts in news videos usually annotates information on where, when, and who of the happening events [1]. Texts in sport video tell us some important events. This has inspired a long research on detecting texts in images and videos [1-8]. These works are mainly based on the following properties of text:

- 1) Dense intensity variety (or gradient);
- 2) Contrast between text and its background;
- 3) Structural information;
- 4) Texture property.

In this paper, we categorize the existing approaches as: Text detection using 1) edge (gradient) feature [2][3][4], 2) using connected component analysis [5], 3) using texture feature [1][6], 4) using temporal features [7][8]. In the first kind of algorithm, “horizontal rectangle structure of clustered sharp edges” [3] will be considered or classified as text blocks. Although these approaches can produce a high recall rate, they produce many false alarms because background may also have edges (gradient) like text. Jain’s work [5] is representative one of second kinds of algorithms. He first decomposes an image into multi-color map and then selects foreground pixel by analyzing connect components of same color. Finally structure properties are adopted to identify text. In the third kind of algorithm, researchers considered text as a kind of texture and use wavelet domain texture feature [6], DCT features [1], etc. to discriminate text with the rest of world.

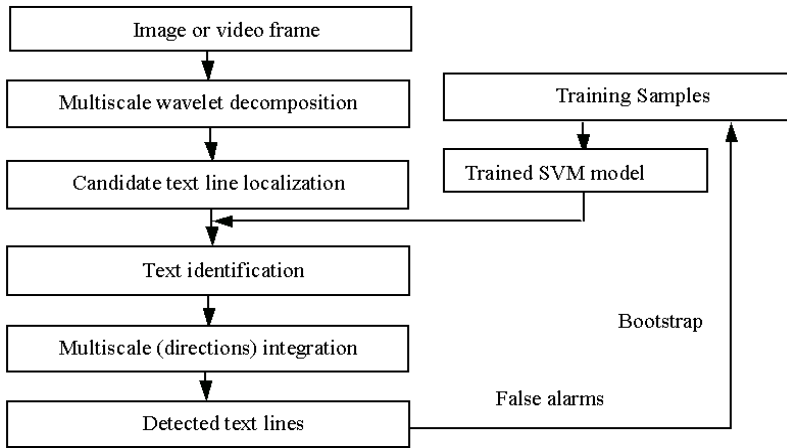


Fig. 1. Flow chart of the proposed algorithm

Video temporal information is also employed for text detection [7][8]. In this paper, we propose a new text detection algorithm in coarse-to-fine framework by integrating properties of text in spatial domain. Property 1), 2) is employed to getting candidate text regions. Property 3) is adopted to separate regions into text lines. Property 4) is used to discriminate text/ nontext . We combine four kinds of features to capture the text’s texture property. A SVM classifier, which has the good generalization ability, is employed to classify text with nontext. The algorithm has a high speed by avoiding classifying the image block by block Fig.1. is the flow chart of proposed algorithm.

In the rest of the paper, coarse detection is presented in section 2, fine detection in section 3. Experimental result is presented in section 4.

2 Text Coarse Detection

We take horizontal text line as an example to state the algorithm. Vertical and skewed text lines can be detected by a rotated the region growing template.

2.1 Image Decomposition with Multiscale Wavelets

Daubichie4 wavelet is selected for represent images for its good location performance [10] which is computed by applying filters on the image as

$$\begin{aligned}
 L_n(b_i, b_j) &= [H_x * [H_y * L_{n-1}]_{\downarrow 1,2}]_{\downarrow 1,2}(b_i, b_j) \\
 D_{n1}(b_i, b_j) &= [H_x * [G_y * L_{n-1}]_{\downarrow 1,2}]_{\downarrow 1,2}(b_i, b_j) \\
 D_{n2}(b_i, b_j) &= [G_x * [H_y * L_{n-1}]_{\downarrow 1,2}]_{\downarrow 1,2}(b_i, b_j) \\
 D_{n3}(b_i, b_j) &= [G_x * [G_y * L_{n-1}]_{\downarrow 1,2}]_{\downarrow 1,2}(b_i, b_j)
 \end{aligned} \tag{1}$$



Fig. 2. Texts in two font-sizes and their wavelet decomposition in two levels

where $*$ denotes the convolution operator, $(\downarrow_{1,2})$ subsampling along rows (columns) and L_0 is the original image, H and G are high and low bandpass filters respectively. b_i, b_j are the locations in two directions, $(L_n, D_{nk})_{k=i,2,3n=1,2,\dots,l}$ is the multiscale representation of l depth of the image. L_n is the low resolution image in scale n , D_{nk} is the wavelet response obtained by filtering containing the intensity variety in scale n that can be adopted to detect text in different font-size. Fig.2 shows that texts with different sizes are emphasized in different scales. It also implies that we can extract text in different sizes from the translated image.

2.2 Candidate Text Region Detection

Considering intensity variety (property 1) around the text pixels, the wavelet coefficients around the pixels must have large values (as shown in Fig.3b). We define the wavelet energy feature (WE) of a pixel at (i, j) in scale n by integrating the wavelet coefficients in the three high frequency subbands (LH, HL and HH subbands) as

$$E_n(b_i, b_j) = \left(\sum_{k=1}^3 [D_{nk}(b_i, b_j)]^2 \right)^{1/2} \quad (2)$$

A pixel will be a candidate text pixel in level n if its WE is larger than a threshold, which is described as

$$C_n(i, j) = \begin{cases} 1 & \text{if } E_n(b_i, b_j) > T_c \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $C_n(i, j)$ is the map of candidate text pixels in level n (Fig.3b, in the map, candidate pixels in first level are projected into original image). T_c is a threshold whose value should ensure that all of the text pixels can be perceived by human be discriminate as candidate pixels. Value of 30.0 is selected for T_c .

A text region is made of a “cluster” of text pixels. Isolated pixels are often noises. Previous researchers [2][9] use morphological algorithm to connect text pixels into text regions. In the operation, all of the pixels near to each other will be connected despite whether they are a “cluster” of pixels or not (Fig.3c). In this paper, “density-based” region growing is proposed to fulfill this task.

A pixel P will be a seed pixel if the percent of candidate pixels in its neighborhood is larger than a threshold T_D . The neighborhood is a 16×10 template



Fig. 3. (a) Original image, (b) Candidate pixels, (c) Candidate text regions by morphological “close” operation and (d) by “density-based” region grow.

in all of the levels. To find text lines in non-horizontal directions, we rotate the template 30 degree each time and getting 6 templates. T_D is set as 0.35 in our experiments. We define that pixel P' is density-connected with pixel P if P' is in the neighborhood of P and P is a seed pixel. By this definition, we can describe region growing method as follows:

- 1) Search the unlabeled pixels to find a seed pixel.
- 2) If a seed pixel P is found, a new region is created. We iteratively collect unlabeled pixels density-connected with P , and label these pixels by same label.
- 3) If there are still seed pixels goto 2).
- 4) Label founds region as text regions.

Some of the detected text regions contain multi-line texts. A projection profile [4] is employed to separate these regions into single-line text. A text line will be discarded if its width/height is smaller than 2.0 or height smaller than 8 pixels.

3 Text Fine Detection

Things taking on violent gradient may be detected as text, especially general textures. In the followings, we will use texture features to reduce false alarms.

3.1 Feature Extraction

Text's texture property such as regularity, directionality is weak, only that it contains strokes forming a text line. We combine four kinds of features to capture the property.

1) *Waveletmoment features* : Text and nontext have different intensity variance and spatial distributions. Wavelet mean and central moments features can reflect these differences. For a $M \times N$ text line , the mean (m), second-order (μ_2) and third order (μ_3) central moments are calculated as

$$\begin{aligned}
 m(T) &= \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} T(i, j) \\
 \mu_2(T) &= \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (T(i, j) - m(T))^2 \\
 \mu_3(T) &= \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (T(i, j) - m(T))^3
 \end{aligned} \tag{4}$$

The features in (4) are calculated in the high frequency subbands (HL, LH and HH subbands) in the level where the candidate text line be found. There are 9 features.

2) *Wavelethistogramfeature* : Histogram is the effective feature to represent the first order distribution of signatures. In this paper we use Wavelet energy histogram (WEH) and Wavelet direction histogram (WDH) to represent the energy and direction distribution of a text line. WEH is equally quantized into 16 dimensions and WDH 4 dimensions. For a text line, the bins at the front and finality of WEH should be large while bins in the middle parts of the WEH should be small. This is caused by the contrast between text and its background. WDH includes horizontal, vertical, dialog, and non-direction bins. We say that a candidate pixel has a horizontal (vertical, dialog) direction when $D_{n1}(D_{n2}, D_{n3})$ is the largest one among D_{nk} . All non-candidate pixels will be non-direction pixels.

3) *Waveletcooccurrencefeatures* : The element (i, j) of the cooccurrence matrix $C(d, \theta)$ is defined as the joint probability a wavelet coefficient $D_{nk} = i$ cooccurs with a coefficient $D_{nk} = j$ on a distance d in a direction θ . The features are energy, entropy, inertia, local homogeneity and correlation of the cooccurrence matrixes as

$$\begin{aligned}
 \text{Energy} : E(d, \theta) &= \sum_{i,j} C^2(d, \theta) \\
 \text{Entropy} : H(d, \theta) &= \sum_{i,j} C(d, \theta) \log C(d, \theta) \\
 \text{Inertia} : I(d, \theta) &= \sum_{i,j} (i - j)^2 C(d, \theta) \\
 \text{Localhomogeneity} : L(d, \theta) &= \sum_{i,j} \frac{1}{1+(i-j)^2} C(d, \theta) \\
 \text{Correlation} : R(d, \theta) &= \frac{\sum_{i,j} (i-\mu_x)(j-\mu_y)C(d,\theta)}{\sigma_x\sigma_y} \tag{5}
 \end{aligned}$$

where μ_x, μ_y and σ_x, σ_y are means and variances of $C(d, \theta)$. θ is the direction selected as 0, 45, 90 and 135 degree. d is the cooccurrence distance, which is set to be one pixel. We get totally 80 features in 4 matrixes in 4 subbands.

4) *Crossingcounthistogramfeature* : The above features do not consider the periodicity of text along the text line. We capture this periodicity in the gradient map (Fig.4b) by proposing a new feature. We first project gradient values into one dimension data along the horizontal direction and get gradient projection map (GPM) (Fig.4c). After a Gaussian smoothing (Fig.4d), it can be seen that, there are regularity and periodicity in the GPM. Then, we use the crossing count histogram (CCH) to measure the periodicity property. A crossing count is the number of times the pixel value change from 0 (white) to 1 along a horizontal raster scan line (Fig.4d gives an example). This feature can be used to measure the complexity and regularity of GMP while is not sensitive to the magnitude of the periodicity. Suppose $C_k, K = 1, 2, \dots, N$, is the crossing count on a horizontal scan line k , the crossing count histogram H_k is calculated as

$$H_k = \frac{C_k}{\sum_{l=1}^n C_l} \tag{6}$$

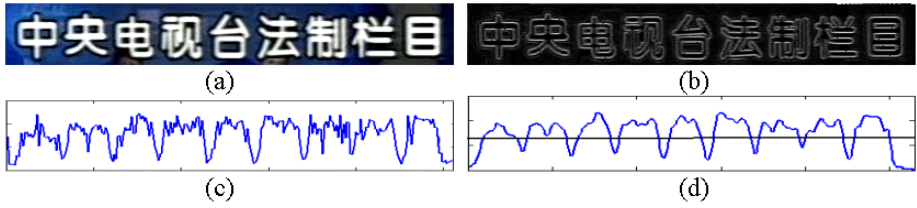


Fig. 4. Crossing count histograms. a) is a text line, b) is the gradient image, c) is the gradient projection and d) is the smooth gradient projection and a scan line.

We equally divide the histogram features into 16 bins. We get totally 141 features, based on which a SVM classifier is employed to classify text with nontext.

3.2 Identification Using SVM

SVM is proposed by Vapnik and have yield excellent results in various two-class classification problems [11] in recent years. The important advantage of the support vector classifier is that it offers a possibility to train generalizable, nonlinear classifiers in high-dimensional spaces using small training set [12]. The SVM classifier uses the structural risk minimization to find the hyperplane that optimally separates two classes of objects. The hyperplane is computed as

$$f(x) = sgn\left(\sum_{i=1}^n y_i \alpha_i K(x, x_i^*) + \alpha_0\right) \tag{7}$$

where *sgn* is a sign function, *K* is a kernel functions, $y_i \in \{-1, 1\}$ is the class label of the data point x . The x_i^* are support vectors and define the separating hyperplane. The parameters α_i are optimized during training. The kernel function used in this paper is a second polynomial $K(x, x_i^*) = (1 + x_i^*)^2$ for its better performance compared with other kernels. The SVM was trained on a dataset consisting of 1,300 text lines and 3,000 non-text lines. After get the trained model, a “bootstrap” [13] process is used to improve the classification performance. Fig.5 is an example of coarse and fine detection results.



Fig. 5. Detection result. (a) is the candidate texts and (b) is the final results



Fig. 6. Final detected text lines

Table 1. Performance comparison of three algorithms

	Recall	False alarm	Speed 25
Our algorithm	93.6%	2.4%	8.0 images/s
Algorithm [3]	91.4%	5.6%	1.5 images/s
Algorithm [6]	94.3%	8.1%	2.2 images/s

4 Experimentation

We select two test sets for experiments One is our test includes 400 images from CCTV video frames and 100 scene images gotten by digital camera. The other is Microsoft test set including 44 images [14]. The image size is 400 x 328 pixels. The test sets consist of a variety of cases, including texts in different sizes, colors and languages, light text on dark background, text on textured background, etc. The algorithm performs robust on a majority of the images as examples in Fig.6.

Recall and *false alarm rate* are used to evaluate as algorithm. *Recall rate* is the percentage of truly text detected as text. *False alarm rate* is the percentage of non-text detected as text. The 93.7%/93.5% and 2.4%/2.4% recall/false alarm rates show the good performance of this algorithm. The great decrease of false alarm rate from the coarse to fine detect has proved the validity of fine detection. The detection rates for Chinese and English text in all of the test sets are 93.4% and 93.7% respectively which shows that proposed algorithm is robust to languages. We compare the proposed algorithm with [6] and [3]. It can be seen (Table 1) that our algorithm has the fastest speed. Both the recall rate and false alarm rate is better than algorithm [3]. Although the recall rate is litter lower than that of algorithm [6], the false alarm (2.4%) is much lower.

5 Conclusion

A new text detection algorithm for image/video frames is proposed, which has high detection speed and low false alarm rate even for text in complex background by using texture features combination and SVM-based classification. Detected result is text line that will benefit the text recognition in future works.

References

1. Y. Zhong, H.J. Zhang, and A. K. Jain: Automatic Caption Localization in Compressed Video. *IEEE Trans. on PAMI*, Vol. 22, No. 4, (2000)385-392
2. V. Wu, R. Manmatha, and E. M. Riseman: Textfinder: An Automatic System to Detect and Recognize Text in Images. *IEEE Trans. on PAMI*, Vol. 20, (1999) 1224-1229.
3. R. Lienhart and A. Wernicke: Localizing and Segmenting Text in Images and Videos. *IEEE Trans. on CSVT*, Vol.12, No.4 (2002)
4. A.K. Jain and B. Yu: Automatic Text Location in Images and Video Frames. *Pattern Recognition*. Vol. 31, No. 12, (1998)2055-2076
5. Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," *Pattern Recognition*. Vol. 28, (1995) 1523-1535
6. H. Li, D. Doermann, and O. Kia: Automatic Text Detection and Tracking in Digital Video. *IEEE Trans. on Image Processing*, Vol. 9, No.1 (2000)
7. X. Tang, X. B. Gao, J. Liu and H. Zhang: Spatial-Temporal Approach for Video Caption Detection and Recognition. *IEEE Trans. Neural Networks*, Vol.13,(2002) 961-971.
8. B. Luo, X.O Tang, J.Z Liu and H. Zhang: Video Caption Detection and Extraction Using Temporal Feature Vector. in: *Int. Conf. on Image Processing*,(2003)
9. D. T. Chen, H. Bourlard, J-P. Thiran: Text Identification in Complex Background Using SVM. *Int. Conf. on CVPR* (2001)
10. S.G Mallat: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. on PAMI*, Vol.11, No.7 (1989)
11. V.N. Vapnik: *The Nature of Statistical Learning Theory*, Springer, (1995)
12. A.K. Jain: Statistical Pattern Recognition: A Review. *IEEE Trans. on PAMI*, Vol.2, No.1, (2001) 4-37
13. K. Sung and T. Poggio: Example-based Learning for View-based Human Face Detection. *Mass, Inst. Technol., Cambridge, MA, A.I. Memo 1521* (1994)
14. X.S. Hua, W.Y. Liu, H.J. Zhang: Automatic Performance Evaluation for Video Text Detection. In: *Int. Conf. on Document Analysis and Recognition* (2001)

Automatic Video Object Tracking Using a Mosaic-Based Background

Young-Kee Jung¹, Kyu-Won Lee², Dong-Min Woo³, and Yo-Sung Ho⁴

¹ Honam University, Gwangju, 506-090, Korea
ykjung@honam.ac.kr

² Daejeon University, Daejeon, 300-716, Korea
kwlee@dju.ac.kr

³ Myongji University, Yongin, 449-728, Korea
dmwoo@mju.ac.kr

⁴ Gwangju Institute of Science and Technology, Gwangju, 500-712, Korea
hoyo@gist.ac.kr

Abstract. In this paper, we propose a panorama-based object tracking scheme for wide-view surveillance systems that can detect and track moving objects with a pan-tilt camera. A dynamic mosaic of the background is progressively integrated in a single image using the camera motion information. For the camera motion estimation, we calculate affine motion parameters for each frame sequentially with respect to its previous frame. The camera motion is robustly estimated on the background by discriminating between background and foreground regions. The modified block-based motion estimation is used to separate the background region. Each moving object is segmented by image subtraction from the mosaic background. The proposed tracking system has demonstrated good performance for several test video sequences.

Keywords: Object Tracking, Object Detection, Image Mosaic.

1 Introduction

In recent years, there have been various research works on image segmentation for object-based coding using the MPEG-4 standard [1,2,3,4,5,6]. MPEG-4 supports a content-based representation of visual objects in the compressed bit-stream domain. Extraction of moving objects plays a key role in such kind of applications. For video conferencing and surveillance, we use a camera system to watch moving objects in the restricted area. If objects move outside the field of view, the camera should pan or tilt such that they always stay within its field of view. In those applications, motion detection and tracking of moving objects play quite important roles.

However, it is difficult to extract video objects and to design general and robust solutions to problems involved. Conventional object extraction and tracking schemes could not be applied to the video sequence taken from an active camera because the moving camera creates image changes due to its own motion. Although a few references [5,6] have addressed the problem of object segmentation

and tracking using an active camera, they cannot segment the moving object in an arbitrary pan and tilt angle.

In this paper, we attempt to resolve these problems. First, we trace a moving object based on the image mosaic background with an active camera. Second, we have focused on how to match images for the general transformations, i.e. for the cases when the camera pans, rotates, and tilts in any directions. An affine model is utilized to generate the image mosaic background. An image mosaic is a panoramic image reconstructed from multiple frames in the video sequence [7,8]. Affine models provide greater flexibility in modelling a global motion, being able to represent rotation, dilation, and shear, as well as translation. Third, the camera motion is robustly estimated on the background by discriminating between background and foreground regions. Therefore, the camera motion estimate is not spoiled by the presence of outliers due to foreground objects whose motions are not representatives of the camera motion.

2 Proposed Tracking Algorithm

As shown in Fig. 1, the proposed tracking system consists of five functional parts: foreground and background region separation, camera motion estimation, mosaic background, object detection and tracking, and control of the pan-tilt camera. After background and foreground regions are identified based on dominant motion estimates, camera motion is then estimated on the background by applying parametric affine motion estimation. The image mosaic background is integrated in a single image. Finally, after we detect and trace the moving object using background, we command the pan-tilt controller to position the moving object at the center of the camera.

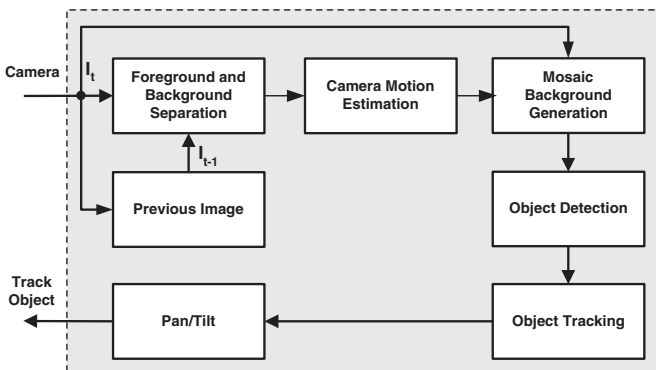


Fig. 1. Proposed Object Tracking Algorithm

2.1 Background and Foreground Separation

The discrimination between background and foreground is based on block-based motion estimation. After the dominant motion is extracted by clustering the block vectors, regions of the dominant motion are identified as background, and otherwise as foreground. This separation has the following two steps: *block-based motion estimation* and *background region extraction*.

Block-based Motion Estimation. In this paper, the modified block-based estimator is used to track changes of the individual block while the global motion estimation step is introduced for deriving a single representative affine motion. Each frame of the 320x240 pixel resolution is divided into non-overlapping of 32x24 pixels. For the block motion estimation, a 9x9 window region with the maximum standard deviation is extracted within each block, as shown in Fig. 2.



Fig. 2. Block-based Motion Estimation (a) the selected 9x9 region for block motion estimation (b) the extracted block vectors

However, in low contrast areas, resulting motion vectors are unreliable. In order to overcome this problem, we apply the activity criterion to filter out unreliable blocks with the lower standard deviation than a certain threshold value. The extracted 9x9 template is correlated in the search region. After we locate the correlation peak, a motion vector is associated with each block. The block motion vector holds the displacement of the block between the current and the previous frames.

Background Region Extraction. In order to extract the background motion, we compute a dominant motion by the following procedure:

- (1) For all block motion vectors, count the number of times that a motion vector is used.

- (2) Obtain the most and second-most popularly used motion vectors.
- (3) Average the two motion vector candidates.

If the motion of the block is similar to the dominant motion, we regard this block as the background block. Finally, foreground or noise blocks are removed.

2.2 Camera Motion Estimation

After the background motion is discriminated from the other motions, the camera motion is estimated from the background. In this way, the camera motion estimate is not disturbed by the presence of foreground objects.

The camera motion is modelled by a parametric affine motion model of six parameters. Once we estimate the six parameters using the least square method from the background motion vectors, we compensate the camera motion through the inverse affine motion transformation.

Let (x, y) be a block position in the previous frame and (x', y') be the position in the current frame. Then, we can represent the motion vector (v_X, v_Y) by

$$\begin{bmatrix} v_X(x, y) \\ v_Y(x, y) \end{bmatrix} = \begin{bmatrix} x' - x \\ y' - y \end{bmatrix} \tag{1}$$

Since we use the affine motion model of six parameters, the motion vector can be expressed as

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_4 & a_5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_3 \\ a_6 \end{bmatrix} \tag{2}$$

In order to estimate six affine motion parameters, we define an error function to be minimized:

$$E(a) = \sum_{i=1}^N \{ [v_X(x_i, y_i) - v_X(x_i, y_i)]^2 + [v_Y(x_i, y_i) - v_Y(x_i, y_i)]^2 \} \tag{3}$$

where N is the number of motion vectors in the same frame.

By substituting Eq. 2 into Eq. 3, we have

$$E(a) = \sum_{i=1}^N \{ [v_X(x_i, y_i) - (a_1x + a_2y + a_3)]^2 + [v_Y(x_i, y_i) - (a_4x + a_5y + a_6)]^2 \} \tag{4}$$

The optimal values of the six parameters are estimated by the least square method. The resulting equation is

$$\sum_{i=1}^N \begin{bmatrix} x_i^2 & x_i y_i & x_i & 0 & 0 & 0 \\ x_i y_i & y_i^2 & y_i & 0 & 0 & 0 \\ x_i & y_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_i^2 & x_i y_i & x_i \\ 0 & 0 & 0 & x_i y_i & y_i^2 & y_i \\ 0 & 0 & 0 & x_i & y_i & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \sum_{i=1}^N \begin{bmatrix} v_X(x_i, y_i)x \\ v_X(x_i, y_i)y \\ v_X(x_i, y_i) \\ v_Y(x_i, y_i)x \\ v_Y(x_i, y_i)y \\ v_Y(x_i, y_i) \end{bmatrix} \tag{5}$$

2.3 Mosaic Background Generation

Once the affine parameters have been calculated, we can warp all the images with respect to the common coordinate system. In order to create the final mosaic image, we map the transformation parameters for each frame into the reference coordinate system by concatenating the transformation matrices. In this paper, we have arbitrarily chosen the first image as the reference, and warped all other images into the first image's coordinate system. Using the camera motion information, a dynamic mosaic of the background is progressively integrated and stored in a single image.

2.4 Object Detection and Tracking

During the camera motion estimation process, the affine motion parameters for each frame are estimated sequentially with respect to its previous frame. From the camera motion information, we can extract the corresponding region of the background mosaic. The moving objects are then segmented by subtracting between the current frame and the corresponding background region.

Once the object region is detected, we can track the object efficiently by predicting the next coordinate from the observed coordinate of the object centroid. We design a 2D token-based tracking scheme using Kalman filtering [9]. The center position and the size of the object are used as the system states to be estimated. After we define the system model and the measurement model, we apply the recursive Kalman filtering algorithm to obtain linear minimum variance (LMV) estimates of motion parameters.



Fig. 3. Initial Image Mosaic Background

3 Experimental Results and Analysis

The proposed tracking system has been tested on several video sequences in indoor environments. Fig. 3 shows the initial mosaic background from 16 images. Four types of sequences are captured, as shown in Fig. 4; right-panning and left-moving person, right-panning and right-moving, left-panning and right-moving person, left-panning and left-moving person.

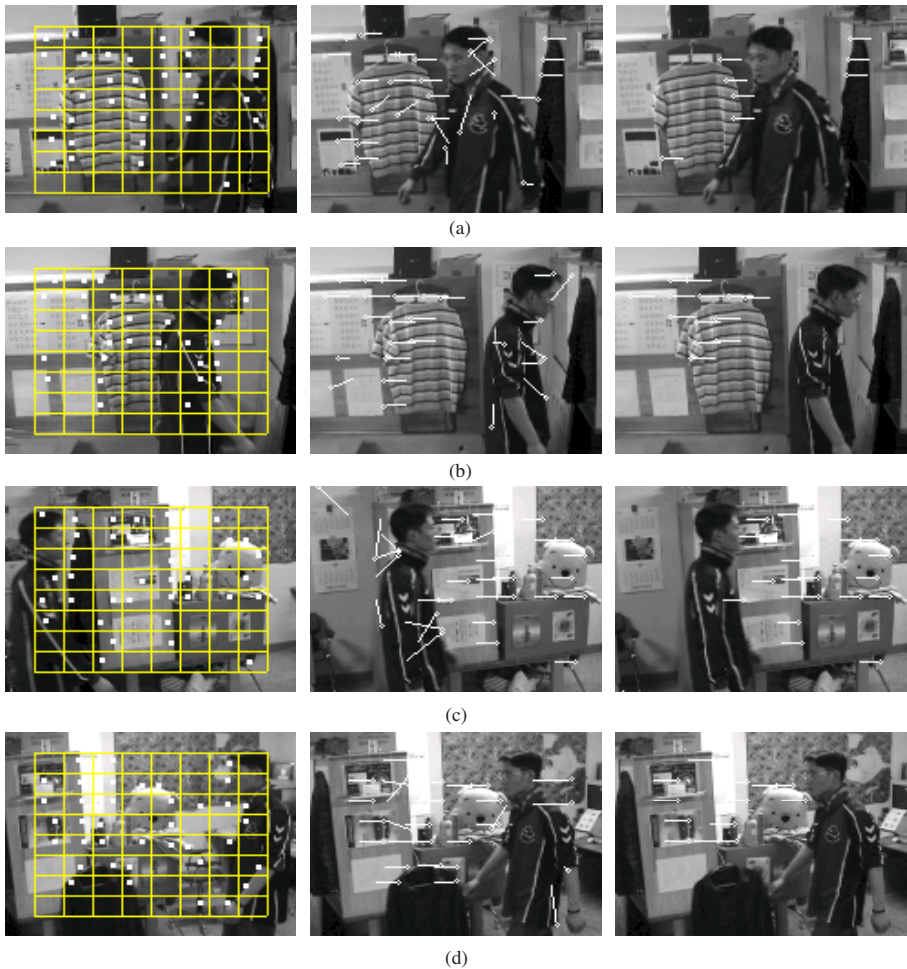


Fig. 4. Background Motion Separation: (a) right-panning and left-moving person, (b) right-panning and right-moving, (c) left-panning and right-moving person, (d) left-panning and left-moving person

In Fig. 4(a), the right panning of camera causes one motion. A moving person occurs the other motion. The background motion is separated by dominant motion vector extraction. The center image of Fig. 4(a) displays the result of block motion vector estimation. The result of background motion separation is represented in the most right image of Fig. 4(a).

Fig. 5(a) shows the mosaic of a room from a sequence of 30 images. Fig. 5(b) shows a current image with one person and the corresponding background from the mosaic. Fig. 5(b) also shows the subtraction image between the current image and the corresponding background image. This subtraction result has some noise blobs due to small errors of camera motion estimation. We utilize a morphological opening operation to remove the noise blobs and the largest blob is chosen to the moving person in Fig. 5(b). The moving person is traced and marked with a white rectangle in Fig. 5(c).

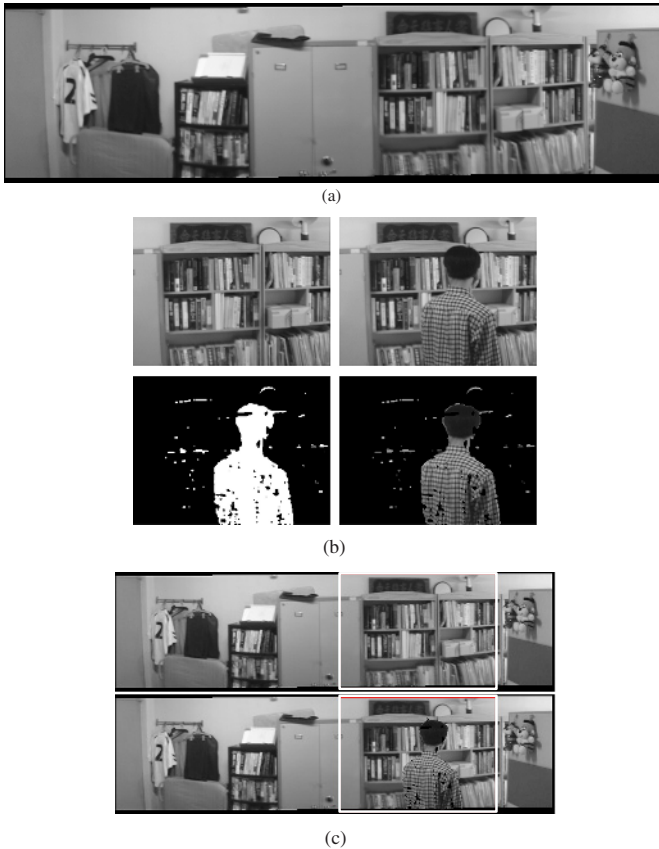


Fig. 5. Tracking Results: (a) mosaic background from a room sequence, (b) extracted object, (c) tracking object

4 Conclusions

In this paper, we propose a new algorithm for moving object tracking using the image mosaic background. We also propose an efficient camera motion estimation algorithm based on background motion to obtain image mosaic integration. In order to build the mosaic, the frames are aligned with respect to a coordinate system and updated. By subtracting the current frame from the corresponding background region, we segment the moving objects. Simulation results demonstrate that the proposed algorithm successfully builds the background mosaic and segments foreground objects.

Acknowledgement. This work was supported in part by grant No. R01-2002-000-00336-0 from the Basic Research Program of the Korea Science & Engineering Foundation, and in part by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation. This work was also supported in part by GIST, in part by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center at GIST, and in part by the Ministry of Education (MOE) through the Brain Korea 21 (BK21) project.

References

1. Gould, K., Shah, M.: The Trajectory Primal Sketch: A Multi-Scale Scheme for Representing Motion Characteristics. IEEE Conf. of CVPR, (1989) 79-85
2. Rouke, O., Badler: Model-based Image Analysis of Human Motion using Constraint Propagation. IEEE Trans. on PAMI, Vol.3, No.4 (1980) 522-537
3. Lee, K.W., Kim, Y.H., Jeon, J., and Park, K.T.: An Algorithm of Moving Object Extraction Under Visual Tracking without Camera Calibration. Proceedings of ICEIC, (1995) 151-154
4. Lipton, A.J., Fujiyoshi, H., and Patil, R.S.: Moving target classification and tracking from real time video. IEEE Workshop on Applications of Computer Vision, (1998) 8-14
5. Ye, Y., Tsotsos, J. K., Bennet, K., and Harley, E.: Tracking a person with pre-recorded image database and a pan, tilt, and zoom camera. Proc. IEEE Workshop on Visual Surveillance, (1998) 10-17
6. Hat, S., Saptharishi, M., and Khosla, P.K.: Motion detection and segmentation using image mosaics. Proc. IEEE Int. Conf. Multimedia and Expo., (2000) 1577-1580
7. Szeliski, R., and Shum, H.: Creating full view panoramic image mosaics and environment maps. Computer Graphics Proceedings, (1997) 251-258
8. Irani, M., Anandan, P., Bergen, J., Kumar, R., and Hsu, S.: Mosaic Representation of video sequences and their applications. Signal Processing: Image Communication, special issue on Image and Video Semantics: Processing, Analysis, and Application, Vol. 8, No. 4, (1996) 673-676
9. Jung, Y.K., Ho, Y.S.: Robust Vehicle Detection and Tracking for Traffic Surveillance. Picture Coding Symposium, (1999) 227-230

Semantic Region Detection in Acoustic Music Signals

Namunu Chinthaka Maddage^{1,2}, Changsheng Xu¹,
Arun Shenoy², and Ye Wang²

¹ Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{maddage,xucs}@i2r.a-star.edu.sg

² School of Computing, National University of Singapore, Singapore 117543
{arunshen,wangye}@comp.nus.edu.sg

Abstract. We propose a novel approach to detect semantic regions (pure vocals, pure instrumental and instrumental mixed vocals) in acoustic music signals. The acoustic music signal is first segmented at the beat level based on our proposed rhythm tracking algorithm. Then for each segment Cepstral coefficients are extracted from the Octave Scale to characterize music content. Finally, a hierarchical classification method is proposed to detect semantic regions. Different from previous methods, our proposed approach fully considers the music knowledge in segmenting and detecting the semantic regions in music signals. Experimental results illustrate that over 80% accuracy is achieved for semantic region detection.

1 Introduction

Music content analysis has become an active research topic in recent years. If the semantic regions in a music signal can be detected, it would be helpful for a better understanding of the music content. For example, in order to build a content-based music retrieval system, the sung vocal line is one of the intrinsic properties in a given music signal.

The singing voice is the oldest music instrument and the human auditory physiology and perceptual apparatus have evolved to a high level of sensitivity to the human voice. After over three decades of extensive research on speech recognition, the technology has matured to the level of practical applications. However, speech recognition techniques have limitations when applied in singing voice identification because speech and singing voice differ significantly in terms of their production and perception by the human ear [15]. Singing voice has more dynamic and complicated characteristics than speech. The dynamic range of the fundamental frequency (F0) contours in singing voice is wider than that in speech and F0 fluctuations in singing voices are larger and more rapid than those in speech [13]. Semantic region detection in music signals is a new direction in music content analysis and it is useful in many tasks, such as singer identification, genre classification, audio source separation, and singing voice removal from the

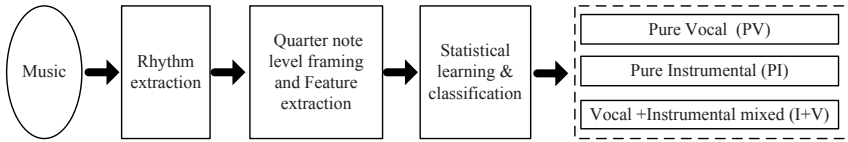


Fig. 1. The Block diagram of the proposed method

music for applications such as Karaoke. Several approaches have been proposed for instrument and singing voice identification.

For instrument identification, Fujinaga [7] trained a K-nearest neighbor classifier with features extracted from 1338 spectral slices representing 23 instruments playing a range of pitches. Eronen and Klapuri [6] proposed a system for music instrument recognition using a wide set of features to model the temporal and spectral characteristics of sounds. For singing voice identification, Berenzweig *et al.* [1] used probabilistic features, calculated using Cepstral coefficients, to train HMM to classify vocal and instrumental music. Kim *et al.* [12] used an IIR filter and an inverse comb filterbank to detect the vocal boundaries. Zhang [17] used a simple threshold which is calculated using energy, average zero crossing, harmonic coefficients and spectral flux features to find the starting point of the vocal part.

Although above mentioned methods have achieved up to 80% of frame level accuracy, their performance is limited due to the fact that music knowledge has not been effectively exploited in existing (mostly bottom-up) methods. We believe that a combination of bottom-up and top-down approaches, which combines the strength of low-level features and high-level music knowledge, can provide us a powerful tool for improved system performance. In this paper, we propose a novel approach, which considers both signal processing and music knowledge, to detect the following semantic regions in acoustical music signals - Pure vocal (PV), Pure instrumental (PI) and Instrumental Mixed Vocals (IMV). The block diagram of proposed method is shown in Fig.1. Since the sung vocal passages follow the changes in the chord pattern, we can apply the following knowledge of chords [8] to the timing information of vocal passages - Chords are more likely to change on beat times than on other positions.

Therefore the music signal is first framed into beat-length segments by extracting metadata in the form of quarter note¹ detection of the music. In section 2, the calculation of beat space length, i.e. quarter note length, is described. Cepstral coefficients are then extracted from the beat space segmented frames based on the *Octave Scale* (section 3). Finally the statistical learning techniques of SVM and GMM are applied to detect semantic regions in musical signals.

¹ This paper uses score-representing terminology as used in [7]. In this formulation, the quarter-note level indicates the temporal basic unit that a human feels in music which in a meter of 4/4, corresponds to a quarter note in scores.

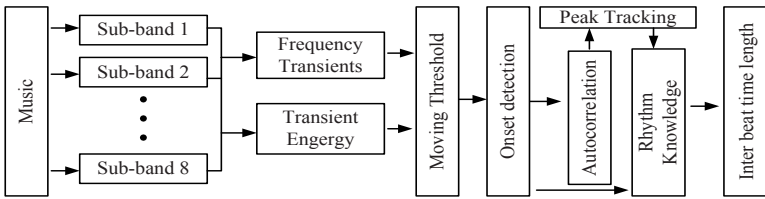


Fig. 2. Rhythm tracking and extraction

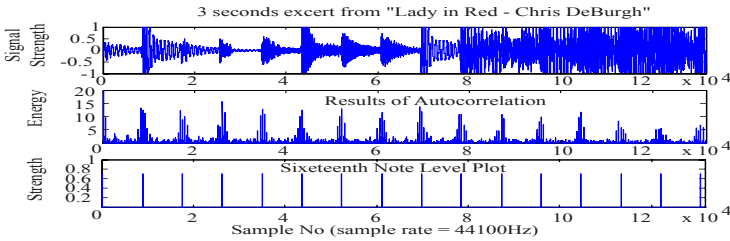


Fig. 3. Three second clip of musical audio signal

2 Rhythm Extraction

Rhythm extraction is important to obtain metadata from music. Rhythm can be perceived as a combination of strong and weak beats [9]. A strong beat usually corresponds to the first and third quarter note in a measure and the weak beat corresponds to the second and fourth quarter note in a measure. If the strong beat constantly alternates with the weak beat, the inter-beat-interval, which is the temporal difference between two successive beats, would correspond to the temporal length of a quarter note. Our proposed rhythm tracking and extraction approach is shown in Fig.2. In our method the beat corresponds to the sequence of equally spaced phenomenal impulses which define the tempo for the music [14]. We assume the meter to be 4/4, this being the most frequent meter of popular songs and the tempo of the input song to be constrained between 30-240 M.M (Mälzel’s Metronome: the number of quarter notes per minute) and almost constant.

The onsets are detected using a sub-band decomposition approach [5]. As the bass drum usually corresponds to the strong beat and the snare drum to the weak beat, we detect the quarter note times, based on a statistical autocorrelation of the bass and snare drum onset times. This can be seen with an example in Figure 3, for an excerpt from the song “*Lady in Red*” by *Chris DeBurgh*. The music is then framed into quarter note spaced segments for further feature extraction which is discussed in the next section.

3 Feature Analysis

We use “Octave” scale instead of “Mel” scale to calculate cepstral coefficients to represent the music content. The Octave scale used previously in [10] handles only 0~8 kHz of the frequency range. In this work, we extend this range to cover the entire audible frequency range, so as to cover a wider range of harmonics. The full frequency range is divided into 8 bands corresponding to the Octaves in music as shown in Tab.1.

Table 1. Sub-band and Octave Scale

Sub Band No	Freq Rang (Hz)	Octave	No. of filters
01	0~128	~B2	6
02	128-256	C3-B3	8
03	256-512	C4-B4	12
04	512-1024	C5-B5	12
05	1024-2048	C6-B6	8
06	2048-4096	C7-B7	8
07	4096-8192	C8-B8	6
08	8192-22050	All higher octaves	4

Then cepstral coefficients [4] are calculated using the responses of the linearly spaced triangular filters in each sub-band according to Eq (1) and Eq(2) where $Y(i)$ is the output of the i^{th} filter, $S(\cdot)$ is the frequency spectrum of the signal frame, N_{cb} is the number of critical band filters, N is the number of frequency sample points, and n is the number of cepstral coefficients. The number of filters in each sub-band is noted in the last column of Tab.1. These cepstral coefficients are called Octave Scale Cepstral Coefficients (OSCC).

$$Y(i) = \sum_{j=m_i}^{n_i} \log |S_i(j)|H_i(j) \tag{1}$$

$$C(n) = \frac{2}{N} \sum_{i=1}^{n_{cb}} Y(i) \cos(k_i \frac{2\pi}{N} n) \tag{2}$$

The useful range of fundamental frequencies of tones produced by music instruments is considerably less than the audible frequency range. The highest tone of the piano has a frequency of 4186 Hz, and this seems to have evolved as a practical upper limit for fundamental frequencies. We have considered the entire audible spectrum to accommodate the harmonics (overtones) of the high tones. The range of fundamental frequency of the voice demanded in classical opera is from ~80-1200 Hz corresponding to the low end of the bass voice and the high end of the soprano voice respectively. It can be seen from Tab.1 that the number of filters are maximum in the bands where the majority of the singing voice is present for better resolution of the signal in that range.

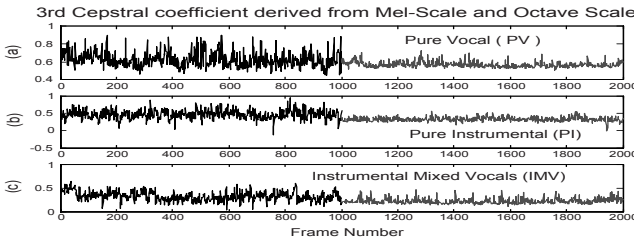


Fig. 4. The 3rd Cepstral coefficient derived from Mel-scale (1~1000 frame in black colored lines) and Octave scale (1001~2000 frames in ash colored lines)

In order to make a comparison, we also extract Cepstral coefficients from the Mel scale. Fig.4 illustrates the deviation of the 3rd Cepstral coefficient derived from both scales for PV, PI & IMV classes. The frame size is quarter note length without overlap. The number of triangular filters used in both scales is 64. It can be seen that the standard deviation is lower for the coefficients derived from the Octave scale, which would make it more robust for our application.

Singular value decomposition (SVD) is applied to find the uncorrelated Cepstral coefficients for both Mel scale and Octave scale. For Mel scale we use the order range from 20~28 coefficients, whereas for the Octave scale the order range of 10~16 coefficients is sufficient according to our investigations.

4 Statistical Learning

To find the semantic region boundaries, we use a two layer hierarchical classification method shown in Fig.4. This has been proven to be more efficient than the single layer multi-class classification method [11]. Support vector machine (SVM) and Gaussian mixture model (GMM) are used as classifiers. SVM has been popular in pattern classification in video and image communities, but has not had much exposure to the audio community. GMM has been used for speaker identification in recent years.

When the classifier is SVM, layers 1 & 2 are modeled with parameter optimized radial basis kernel function [11]. We use SVM Torch II [3] to model the music content with SVMs. The Expectation –Maximization (EM) [2] algorithm is used to estimate the parameters for layer 1 & 2 in GMMs. The GMM can be

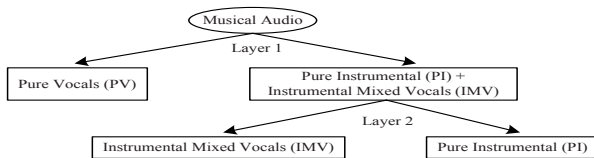


Fig. 5. Hierarchical classification

viewed as a one state HMM. Therefore HTK [16] is used to model different class of music content for GMM classifier.

5 Experimental Results

Our experiments are performed using 15 popular English songs (*Westlife*-5 songs, *Backstreet Boys*-5 songs, *Michel Learns to Rock*-5 songs) and 5 Sri Lankan songs. All music data is sampled from commercial music CDs at a 44.1 kHz sample rate, and 16 bits per sample in stereo.

We first conduct the experiments on our proposed rhythm tracking algorithm to find the quarter note time intervals. The algorithm is run over first 30s, 60s and full length of the song and then computes the average quarter note length. Our system has managed to obtain 95% of average accuracy in the number of beats detected with a ± 20 ms error margin on the quarter note time intervals. The music is then framed into quarter note spaced segments and experiments are conducted for the detection of the class boundaries (PV, PI, & IMV) of the music. 20 songs are used by cross validation where 3/2 songs of each artist are used for training and testing respectively in each turn. We perform three types of experiments:

EXP1 - training and testing songs are divided in to 30ms and 50% overlapping frames.

EXP2 - training songs are divided in to 30ms and 50% overlapping frames and testing songs are framed according to quarter note time interval.

EXP3 - training and testing songs are framed according to quarter note time intervals.

The order of Cepstral coefficients used for both Mel scale and Octave scale in the layer 1 & 2 in both classifiers (SVM & GMM) are shown in Fig.5. 36 and 62 filters are used for calculating cepstral coefficients in both Mel and Octave scale respectively. The number of Gaussian mixtures used for modeling each class in both layers is described in the last column in Table 2.

Table 2. Orders of Cepstral coefficients and no. of GMs employed

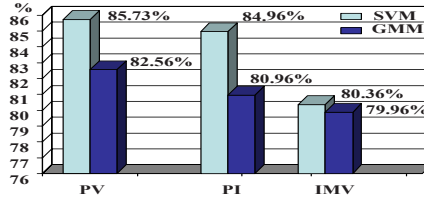
Layers	Mel scale	Octave scale	No of Gaussian Mixtures (GMs)
<i>Layer 1</i>	18	10	24-(PV), 28-(PI+IMV)
<i>Layer 2</i>	24	12	48-(PI), 64-(IMV)

Experimental results (in % accuracy) using SVM learning are tabulated in Tab.2. Mel scale is optimized by having more filters position on lower frequencies [4] because dominant pitches of vocals and musical instruments are in lower frequencies (< 4000 Hz) [15].

Though the training and testing data of PV are small (not many songs have sections of only singing voice “PV”), it is seen that in all experiments (EXP 1~3) the classification accuracy of PV is higher than other two classes. However, when

Table 3. Results of hierarchal classification using SVM (in % accuracy)

Classes Classifier		Mel-scale			Octave scale		
		PV	PI	IMV	PV	PI	IMV
SVM	EXP1	72.35	68.98	64.57	73.76	73.18	68.05
	EXP2	67.34	65.18	64.87	75.22	74.15	73.16
	EXP3	74.96	72.38	70.17	85.73	82.96	80.36

**Fig. 6.** Comparison between SVM and GMM in EXP3

vocals are mixed with instruments, it can be seen that finding vocal boundaries is more difficult than other two classes.

The results of EXP1 demonstrate the higher performance of the Octave scale compared with the Mel scale for 30ms frame size. In EXP2, a slightly better performance can be seen for the Octave scale but not for the Mel scale compared with EXP1. This demonstrates that Cepstral coefficients are sensitive to the frame length and the position of the filters in Mel or Octave scales. EXP3 is seen to achieve the best performance among the EXP 1~3 demonstrating the importance of the inclusion of music knowledge in this application. Furthermore, the better results obtained by use of Octave scale demonstrate its ability to be able to model music signals better than Mel scale for this application.

The results obtained for EXP3 using SVM and GMM are compared in Tab.3. It can be seen that SVM performs better than GMM in identifying the region boundaries. We can thus infer that this implementation of SVM, which is a polynomial learning machine that uses a radial based kernel function, is more efficient than the GMM method that uses probabilistic modeling method using EM algorithm.

6 Conclusion

We have presented a novel approach for the detection of semantic regions in acoustic music signals. A robust rhythm tracking and extraction method is proposed and used to segment music signals at the beat level. This enables us to use musically meaningful inter-beat-interval as the time resolution of our system. A novel feature extraction technique based on the *Octave Scale* is also proposed and illustrates better performance than Mel scale based feature extraction. The Octave scaled Cepstral coefficients are more robust to dynamic changes of the

music compared with the Mel scale. The use of quarter note intervals as the frame length has further improved the performance of our system. We have achieved over 80% accuracy in semantic region classification by using SVM classifier.

References

1. Berenzweig, A.L., Ellis, D.P.W., "Location singing voice segments within music signals" In Proc. IEEE WASPAA, Paltz, New York. Oct (2001).
2. Bilmes, J., "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models". Technical Report ICSI-TR-97-021, University of Berkeley (1998).
3. Collobert, R., Bengio, S., "SVM-Torch: Support Vector Machines for Large-Scale Regression Problems" Journal of Machine Learning Research. Vol 1, (2001), 143-160.
4. Deller, J. R., Hansen, J.H.L., and Proakis, H. J. G., *Discrete-Time Processing of Speech Signals*, IEEE Press (2000).
5. Duxburg, C., Sandler, M., and Davies, M., "A Hybrid Approach to Musical Note Onset Detection", In Proc. of DAFX, Germany (2002).
6. Eronen, A. and Klapuri, A., "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features", In Proc. of ICASSP, Istanbul (2000).
7. Fujinaga, I., "Machine Recognition of Timbre Using Steady-state Tone of Acoustic Musical Instruments", In Proc. of ICMC, (1998) 207-210.
8. Goto, M., "An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds", Journal of New Music Research, Vol.30, No.2 (2001) 159-171.
9. Goto, M., and Muraoka, Y., "A Beat Tracking System for Acoustic Signals of Music", In Proc. of the Second ACM Intl. Conf. on Multimedia, (1994), pp. 365-372.
10. Jiang, D., et al., "Music Type Classification by Spectral Contrast Features", In Proc. of ICME, Switzerland (2002).
11. Maddage, N.C., Changsheng Xu and Wang, Y. "A SVM-Based Classification Approach to Musical Audio" In Proc of ISMIR, Maryland USA (2003).
12. Kim, Y.K. & Brian, W., "Singer Identification in Popular Music Recordings Using Voice Coding Features". In Proc of ISMIR, France (2002).
13. Saitou, T., Unoki, M. and Akagi, M., "Extraction of F0 Dynamic Characteristics and Developments of F0 Control Model in Singing Voice" In Proc. of ICAD, Japan (2002).
14. Scheirer, E.D., "Tempo and Beat Analysis of Acoustic Musical Signals", In JASA, (1998), 103(1).
15. Sundberg, J, *The Science of the Singing Voice*, Northern Illinois University Press, Dekalb, Illinois (1987).
16. Young, S. *et al. The HTK Book*. Version 3.2, (2002).
17. Zhang, T. "Automatic singer identification," In Proc of ICME, Maryland USA (2003).

Audio Classification for Radio Broadcast Indexing: Feature Normalization and Multiple Classifiers Decision

Christine S enac^{1,2} and Eliathamby Ambikairajh²

¹ Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS
118, route de Narbonne, 31062 Toulouse Cedex 04, FRANCE
senac@irit.fr

² School of Electrical Engineering and Telecommunications,
University of New South Wales, Sydney 2052, AUSTRALIA
ambi@ee.unsw.edu.au

Abstract. This paper presents a system that detects the two basic components (speech and music) in the context of radio broadcast indexing. The originality of the approach covers three different points: a differentiated modelling based on Gaussian Mixture Model (GMM), which permits the extraction of speech and music components separately, the normalization of commonly used features and the efficient fusion of classifiers for speech classification which provides a substantial improvement in the presence of strong background music: accuracy of the indexing system goes from [69.2%,94.2%] for the best classifier to [90.25%,98.56%] for the fusion. Evaluation was performed on 12 hours of radio broadcast recorded under various noise conditions, channels and containing diverse speech and music mixtures.

1 Introduction

With the proliferation of audio data from various sources such as television, radio, telephone and the Internet, there has been a large demand for the monitoring of multimedia applications such as audio and broadcast news indexing. In this context, content-based classification of audio data is an important problem and though the classification of audio into pure classes, such as speech, music and silence, is widely studied, classification of mixed audio data is still considered a difficult problem.

In most studies, the first partitioning in audio indexing consists of speech/music discrimination [1,2]. Initially, two main tendencies were observed: while the musical community gave greater importance to features which increase a binary discrimination [3], the automatic speech processing community preferred cepstral parameters. Since then, many approaches have been proposed to classify audio using a different fusion of audio features [4,5] and classifiers such as the GMM, k-Nearest-Neighbours, Neural Networks and Support Vector Machines (SVM).

In this paper, we describe a system that can detect the two basic components (speech and music) in the context of radio broadcast indexing. In a previous paper [6], we used a differentiated modelling approach permitting to exploit the structural difference between speech and music. The originality of this approach is that music and speech are not considered as two concurrent classes because *two classification systems are independently defined* to better characterize and discriminate each component. In particular, two different feature spaces are defined - spectral and cepstral features - while the modelling is based on GMM. Two other original points of our study are to firstly *normalize features*, by the warping feature, to improve the two classifiers, and secondly to propose a robust *fusion algorithm*, for which no training phase and threshold are necessary, for speech classification.

This paper is divided into two main parts: the first section describes the system architecture including the different front-end processings and the classification methods. The second section contains the experimental results including the database description and evaluation for the different classifiers.

2 The System Architecture

The indexing system is constructed around a modular architecture. The extraction of speech and music parts being made in a separate way, the system is divided in two sub-systems. Each one consists of 3 modules: the acoustic front-end processing, the classification module and a smoothing module.

2.1 Acoustic Front-End Processing

For the speech detection system, we *compared two classifiers* and *fused* both of them, based on different front-end processings. Likewise, we compared two classifiers based on different front-end processings for the music detection system (Figure 1).

First Front-End Processing for Speech/NoSpeech Detection. In this system, *Mel-Frequency Cepstral Coefficients* are derived from mel-spaced filter banks log-energies. There are 24 triangular filter banks spanning the band limited region with 8 MFCCs derived from these, using 10ms frames and a 30ms window. Energy and corresponding delta coefficients, calculated by linear regression derivative over a 5 frames window, are also appended. Then cepstral mean subtraction (CMS) is applied.

First Front-End Processing for Music/NoMusic Detection. The parameters are spectral based and extracted from the signal every 10ms using a 30ms window. Each frame of 10ms is pre emphasized and we apply a Hamming window before computing a Fast Fourier Transform. A spectral filtering produces 28 channels through a linear scale. Each acoustic vector is composed of energy and of 28 *spectral coefficients* covering the frequency range [100 Hz-8 kHz].

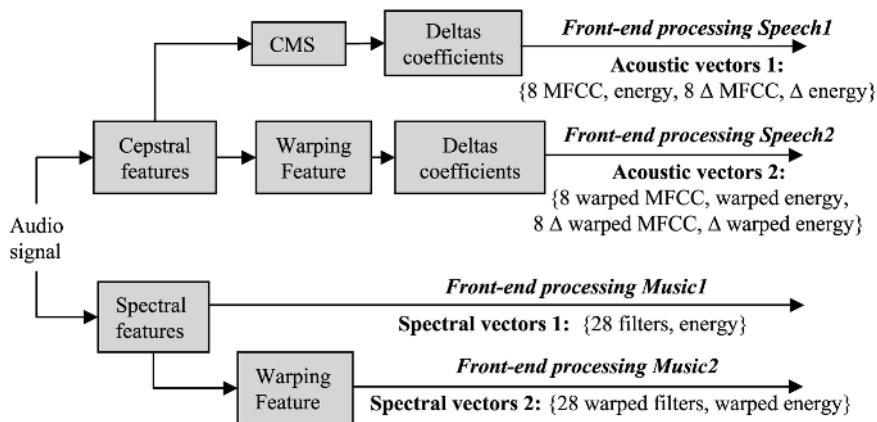


Fig. 1. The different front-end processings used for speech and music classification

Second Front-End Processing for Speech/NoSpeech and Music/No-Music Detection. Here *warping feature* is used in the aim of providing a feature enhancement. Feature warping [7] was introduced for speakers verification applications where there is a need to extract information from speech that is speaker specific and robust to noise and various channel and transducers effects.

Feature warping consists of mapping the observed feature distribution to a normal distribution over a sliding window, the different features being processed in parallel streams. Let r_t the rank of a feature within a N sample window centred around time t , its warped value w_t is estimated by solving numerically the following equation:

$$\frac{rt - 1/2}{N} = \int_{-\infty}^{wt} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{1}$$

The second front-end processing for Speech/NoSpeech detection consisted in warping 8 MFCC and energy to which we appended the deltas. Likewise, for music detection, the second front-end processing consisted in warping 28 filter outputs and energy.

2.2 Class Modelling

Modelling is based on *GMMs* with diagonal covariance matrix. So the training consists in an initialization step using Vector Quantization based on the algorithm of Lloyd followed by an optimization step made by the classic algorithm E-M. For each method, we tried different numbers (from 32 to 1024) of Gaussian mixtures. After experiments, the number of Gaussian mixtures has been fixed to 128 for all the models.

2.3 Classification

For each individual classifier, we chose to model the Class and NoClass by GMM. The classification is achieved by testing the candidate audio frame (of 10ms) against the Class model and the NoClass model.

For a given test frame F and a target model M_{Class} , the decision score $S(F, M_{Class})$ is the log-likelihood (LL) ratio,

$$S(F, M_{Class}) = LL(F, M_{Class}) - LL(F, M_{NoClass}) \quad (2)$$

The frame is labelled with the Class (respectively NoClass) name if $S(F, M_{Class})$ is positive (respectively negative).

It has been demonstrated that the application of *decision fusion algorithms* to combine multiple individual classifiers can enhance the recognition performance and the robustness of a classification system [8] whatever it may be.

The reason for this enhancement lies in the way multiple individual classifiers combination approaches can exploit additional information extracted from the data sets making efficient usage of the complementary information present among the various participating classifiers.

There are various levels of fusion for combining two (or more) classifiers:

- *Fusion at the matching level* where each classifier provides a matching score indicating the proximity of the feature vector with the appropriate models. These scores are combined to point out the recognized Class;
- *Fusion at the decision level* where each classifier individually classifies in Class or NoClass. A further decision is taken according the different outputs of each classifier.

A variety of fusion schemes have been described in the literature, mainly in biometrics. These include majority voting, sum and product rules [9], SVM [10], decision tree [11], Bayesian methods, consensus decision [8], logistic regression [12], ...

For our application, we fused the two Speech classifiers. The fusion took place on the same time at the matching level and at the decision level. The decision fusion algorithm we used is described in the next section.

2.4 Smoothing

Following this frame based classification a phase of merging allows to concatenate neighbouring frames having obtained the same index during the classification. Then insignificant size segments (i.e. < 20ms) are merged and only significant speech (respectively music) segments are kept [6].

3 Experimental Results

3.1 Corpus

Our audio indexing evaluation corpus comprises approximately 12 hours of radio broadcasts from the multilingual Radio France International broadcast, sampled

at 16 kHz. *The recorded programs are very diverse* in terms of speech and music content, speaking styles, speakers, noise conditions and channels.

Three different types were employed:

- *Type I comprises broadcast news* (4 hours), with 17 different speakers (adult male, adult female and children) containing mainly interviews and reports from correspondents. Each broadcast begins with a signature announcement followed by the signature music. The background music is relatively quiet;
- *Type II comprises broadcast interviews* (4 hours) in many channel conditions, and the background music is much more annoying than in type I;
- *Type III (4 hours) comprises diverse musical programs interspersed with interviews under adverse conditions*: different channels, outdoor recordings, crowd noise, phone calls, barely audible speech, very audible background music, songs and ‘rap’ music.

For type I, we used 2 hours for training and 2 hours for testing (3 speakers from the 17 appeared for both training and testing sub corpora). For type II, we used 2 hours for training and 2 hours for testing (the speakers and type of music were different between training and testing). *Type III was tested without training.*

The entire corpus was manually transcribed in ‘speech’ ‘no-speech’, ‘music’ and ‘no-music’ using the labelling tool Transcriber [13].

Training for the speech (respectively no-speech) model used each part of the files containing speech (respectively no speech) whatever the presence or no of music or noise in the background. The same method was applied to train music and no-music models.

3.2 Music/NoMusic Classification

Evaluation of the automatic indexing (Table 1) was made by comparison with the manual labelling. The *accuracy* was computed as follows:

$$\text{Accuracy} = (\text{length}_{\text{testcorpus}} - \text{length}_{\text{insertions}} - \text{length}_{\text{substitutions}}) / \text{length}_{\text{testcorpus}}$$

For broadcast news (Type I), the results show that feature warping is similar to the spectral features. *A significant improvement is seen for feature warping in presence of audio documents recorded in very bad or diverse conditions* (Types II and III). Corpus III, which wasn’t trained, comprises a lot of errors occurring during songs and rap music that would necessitate a special class.

Table 1. Accuracy in % of Music indexing using different classifiers (128 Gaussians)

Features	Type I	Type II	Type III
28 filter outputs and energy	90.23	61.2	52.56
28 warped filter outputs and warped energy	90.44	80.97	71.93

Table 2. Accuracy in % of Speech indexing using different classifiers (128 Gaussians)

Features	Type I	Type II	Type III
Classifier Speech1 (Acoustic vectors 1)	94.20	75.79	56.77
Classifier Speech2 (Acoustic vectors 2)	89.75	77.20	69.20
Fusion of the two above classifiers	98.56	92.26	90.25
Corresponding Kappa statistic	0.28	0.29	0.21

Table 3. The fusion algorithm for speech component classification. The notations are S for speech and NS for non-speech

If Decision Speech1	& Decision Speech2	& Condition	then Fusion Decision
S	S	-	S
NS	NS	-	NS
S	NS	$(S_1 \geq -S_{2norm})$	S
S	NS	$(S_1 < -S_{2norm})$	NS
NS	S	$(S_{2norm} \geq -S_1)$	S
NS	S	$(S_{2norm} < -S_1)$	NS

3.3 Agreement Between Classifiers

We tried to fuse the decisions obtained by these two classifiers but the accuracy obtained after the fusion was lower than those obtained with the warped features.

This is due by the fact that *the two classifiers have a too big agreement* which can inhibit the gains obtained regardless of the method used to combine those.

The level of agreement between different classifiers can be assessed based on statistical measures and especially on the *Kappa statistic* (see [14] for the calculation). A value of kappa below 0.40 is considered to represent poor agreement beyond chance (so the fusion will certainly improve the accuracy), value between 0.40 and 0.75 indicate fair agreement and value beyond 0.75 indicate excellent agreement [15].

The values of the kappa statistic between the two music classifiers for the different audio documents are: 0.53 for type I, 0.49 for type II and 0.48 for type III. These values confirm the too high agreement between the two music classifiers.

3.4 Speech/NoSpeech Classification

The performances of the two types of features for speech classification are shown in Table 2. MFCCs perform better in the presence of broadcast news documents under clean condition (Type I). In the presence of strong background music not represented in the models, a substantial improvement is seen for warped MFCCs (Type II and III). It appears from Table 2 that following the value of the Kappa statistic, the two classifiers contain relatively orthogonal information, and their fusion significantly improves classification results.

Fusion Algorithm. The fusion algorithm (Table 3) takes place at the same time at the decision level (if the two classifiers are in agreement) and at the matching score level (otherwise). *This algorithm does not need training phase or threshold.*

The notations used are: S_1 for scores for classifier Speech1 using feature vectors 1; S_2 for scores for classifier Speech2 using feature vectors 2; S_{2norm} for normalized scores for classifier Speech2 using feature vectors 2.

It's important in fusion at the matching score level to normalize the scores obtained from the different classifiers. Here, normalization involves mapping the scores S_1 and S_2 obtained by the two classifiers into a *common domain* [16]. We chose S_1 scores domain as common domain and scaled the S_2 scores into this domain.

4 Conclusion

We presented a speech/music classification in the context of radio broadcast indexing. A differentiated modelling approach was implemented from GMM. For each component, we processed two kinds of features: the commonly used features (derived from a cepstral analysis for speech and from a spectral one for music) and these same features warped.

We found that feature warping outperformed the basic features for music classification whereas for speech classification they outperformed the basic features in the noisiest conditions (particularly with strong music in the background) but degraded in good conditions.

So we proposed a fusion algorithm which does not need training phase or threshold and we got a substantial improvement whatever the conditions of recording: the rate error was three times smaller than that of either the warped features or the traditional features alone.

References

1. El-Maleh, K., Klein, M., Petrucci, G., Kabal, P., McGill, P.: Speech/music discrimination for multimedia applications. In: Proc. IEEE ICASSP. (2000)
2. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: Proc. IEEE ICASSP. (2003)
3. Rossignol, S., Rodet, X., Soumagne, J., Collette, J., Depalle, P.: Automatic characterization of musical feature extraction and temporal segmentation. *Journal of New Music Research* **28** (1999)
4. Piquier, J., Rouas, J., André-Obrecht, R.: Robust speech/music classification in audio documents. In: Proc. IEEE ICASSP. (2003)
5. Ghaemmaghami, S.: Audio segmentation and classification based on a selective analysis scheme. In: Proc. IEEE MMC. (2004)
6. Razik, J., Sénac, C., Fohr, D., Mella, O., Parlangeau, N.: Comparison of two speech/music segmentations systems for audio indexing on the web. In: Proc. SCI. (2003)

7. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. of 'A Speaker Odyssey'. (2001)
8. Faifhurst, M., Rahman, A.: Enhancing consensus in multiple expert decision fusion. In: IEEE VISIP. Volume 147. (2000)
9. Kwon, O., Lee, T.: Optimizing speech/non-speech classifier design using adaboost. In: Proc. IEEE ICASSP. (2003)
10. Lu, L., Li, S., Zhang, H.: Content based audio segmentation using support vector machines. In: Proc. IEEE ICME. (2001)
11. Ross, A., Jain, A.: Information fusion in biometrics. *Pattern Recognition Letters* 24 (2003)
12. Verlinde, P., Chollet, G.: Comparing decision fusion paradigms using k-nn based classifiers, decision trees and logistic regression in a multi-modal identity verification application. In: Proc. AVPA. (1999)
13. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication* **33** (2000)
14. Petralos, M., Benediktsson, J.: The effect of classifier agreement on the accuracy of the combined classifier in decision level fusion. *IEEE Trans. on Geosciences and Remote Sensing* **39** (2001)
15. Fleiss, J.: *Statistical methods for rates and proportions*. 2nd edn. Wiley (1981)

Dominant Feature Vectors Based Audio Similarity Measure

Jing Gu^{*1}, Lie Lu², Rui Cai³, Hong-Jiang Zhang², and Jian Yang¹

¹ Dept. of Electronic Engineering, Tsinghua Univ., Beijing, 100084, China

² Microsoft Research Asia, Beijing, 100080, China

³ Dept. of Computer Science and Technology, Tsinghua Univ., Beijing, 100084, China

Abstract. This paper presents an approach to extracting dominant feature vectors from an individual audio clip and then proposes a new similarity measure based on the dominant feature vectors. Instead of using the mean and standard deviation of frame features in most conventional methods, the most salient characteristics of an audio clip are represented in the form of several dominant feature vectors. These dominant feature vectors give a better description of the fundamental properties of an audio clip, especially when frame features change a lot along the time line. Evaluations on a content-based audio retrieval system indicate an obvious improvement after using the proposed similarity measure, compared with some other conventional methods.

1 Introduction

A fundamental step of the audio content analysis is similarity measure based on the various features. In a general way, one extracts several features, including temporal and spectral features, from an audio clip (usually last for several seconds), then, the similarity measure between two audio clips is based on the feature vectors constructed by those features. In most cases, the clip is too long, so that it is usually divided into several frames to catch the short-time property. The features are extracted from each frame and their mean and standard deviation are calculated to form the final feature vector of the audio clip. Such clip-based statistical features have proved their effectiveness in many previous works such as [1][2]. However, the frame features of an audio clip usually change much along the time line, that is, there usually exist more than one salient characteristic in the clip. Only the mean and standard deviation of frame features can not give an accurate presentation of the property of such audio clips [3].

For example, Fig. 1 (a) illustrates the spectrogram of a sound of applause. The sound shows periodicity and there are approximately two salient characteristics in the clip, one is for sound period and the other for silence period. Fig. 1 (b) gives another example of a sound made by a jet plane flying over the heads. The spectrogram shows an obvious spectral change due to the “Doppler

* This work was performed when the first author was a visiting student in Media Computing Group, Microsoft Research Asia

effect". Therefore, the characteristics are quite different between two halves of the sound clip. Thus, two or even more salient characteristics are needed to describe this sound. In both cases, neither the mean nor the standard deviation of the frame features gives accurate description of different salient characteristics. Much information will be lost if one just uses them and thus will lead to inaccurate similarity measure in some cases. Therefore, it would be better to find a way to directly represent the most distinct characteristics of an audio clip.

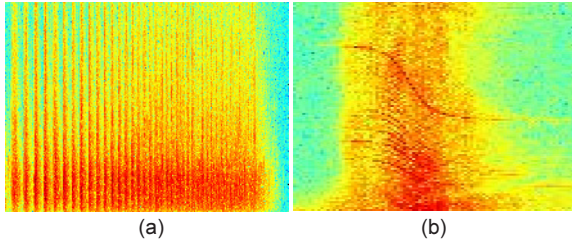


Fig. 1. Sound spectrograms of (a) *applause* and (b) *jet plane*.

In this paper, we propose a new approach to obtain salient characteristics of an audio clip by employing subspace decomposition on the set of frame-based features. The obtained dominant feature vectors describe the most salient characteristics of an audio clip, and can represent the audio clip better than the mean and standard deviation. The number of the dominant feature vectors needed to describe an audio clip is related to the feature variation in the corresponding clip. For example, it needs more dominant feature vectors if the characteristics of the clip change significantly; and less are needed when the characteristics keep stationary.

Suppose an audio clip has several salient characteristics or several dominant feature vectors, it is intuitive that further analysis should be implemented based on these dominant feature vectors. According to this, we further propose a new similarity measure between two audio clips. The proposed similarity measure considers the similarity between each pair of dominant feature vectors, and thus keeps the richness of sound property but reduces the noise.

The rest of the paper is organized as follows. Section 2 presents the detail approach to obtain the dominant feature vectors from an audio clip. Section 3 proposes a new similarity measure between clips and compares it with other conventional similarity measures. The evaluation of the proposed approach is given in the Section 4.

2 Dominant Feature Vectors

In this section, we extract the dominant feature vectors from an audio clip, in order to represent the multiple salient characteristics of the clip.

Assuming that an audio clip is divided into N frames, from each of which an n -dimensional feature vector is extracted and normalized to be zero mean and unit variance over the whole database. The normalized time-varying feature vectors of a clip can be represented by $X = (x_1, x_2, \dots, x_N)$, which is an n -by- N matrix and x_i ($i = 1, 2, \dots, N$) is the feature vector of the i^{th} frame. Now we want to obtain several dominant feature vectors which may give a good description of the audio clip, especially when several salient characteristics exist. Fortunately, we can achieve this object by employing eigen-decomposition on the covariance matrix of the frame based feature vector.

The n -by- n covariance matrix can be estimated as following:

$$C = \frac{1}{N} X X^T \tag{1}$$

By eigen-decomposition, the covariance matrix is decomposed as:

$$C = Q^T \Lambda Q = \sum_{i=1}^n \lambda_i q_i q_i^T \tag{2}$$

where $Q = (q_1, q_2, \dots, q_n)$ is an orthogonal matrix, q_i ($i = 1, 2, \dots, n$) is the eigenvector of C , and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is a diagonal matrix of non-negative real eigen-values ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.

The main idea of this eigen-decomposition is to find the vectors which best describe the characteristics of the set of frame feature vectors, within the space spanned by them. Generally, the eigen-vectors associated with large eigen-values represent the dominant information and can be considered as dominant feature vectors, while those eigen-vectors with small eigen-values have little contribution and can be considered as introduced by noise. Therefore the eigen-vectors associated with large eigen-values are the dominant feature vectors we want to obtain. The corresponding eigen-values can be considered as the importance or the contribution of the dominant feature vectors.

The number of dominant feature vectors needed to represent an audio clip is related to the characteristic variation of a clip. It needs more dominant feature vectors if the characteristics of the clip change much. Considering the eigen-values represent the contribution of the corresponding eigen-vectors, a general way of choosing the number of dominant feature vectors is as follows:

$$m = \arg \min_k \left\{ \sum_{i=1}^k \lambda_i / \sum_{i=1}^n \lambda_i \geq \eta \right\} \tag{3}$$

where m is the number and the threshold $\eta \in (0, 1)$.

The m dominant feature vectors actually span an m -dimensional subspace which is the best approximation of the original n -dimensional eigen-space, suppressing the effect of noise. We call the m -dimensional subspace “*signal subspace*”. The remaining $(n - m)$ -dimensional subspace is called “*noise subspace*”. Correspondingly, the noise-reduced covariance matrix can be represented by

$$C = \sum_{i=1}^m \lambda_i q_i q_i^T \tag{4}$$

It should be noted that our approach to dominant feature vector extraction is totally different with traditional PCA applications. PCA is traditionally used to remove the noisy feature dimensions. However, our approach is used to remove the noisy feature vectors so that the dimension of each feature vector is not decreased. Moreover, dominate feature vectors is performed on an individual audio clip and form a "signal subspace" which represents the most salient characteristics of the corresponding audio clip, while PCA usually is based on the whole database to find the principle feature components and then map each audio clip into one vector in the principle feature space.

3 Dominant Feature Vector Based Similarity Measure

Based on the extracted dominant feature vectors, a new similarity measure is correspondingly proposed in this section. The characteristics of the measure are discussed and the comparisons with other conventional methods are also given.

3.1 Similarity Measure Definition

Consider two audio clips which contain m_1 and m_2 dominant feature vectors respectively, their i^{th} and j^{th} dominant feature vectors is denoted as q_i and p_j , and the corresponding eigen-value is λ_i and σ_j . To measure the similarity between these two audio clips, the similarity between q_i and p_j is firstly considered, which is usually defined as their inner-product:

$$s_{i,j} = \frac{\|q_i^T p_j\|^2}{\|q_i\|^2 \|p_j\|^2} = \|q_i^T p_j\|^2 \tag{5}$$

Since different dominant feature vector has different importance, which is determined by the corresponding eigen-values, in representing an audio clip, they should have different contributions to the audio similarity measure. That is to say, the dominant feature vectors with large eigen-values should contribute more in similarity measuring between two audio clips. Thus, the similarity of two audio clips is defined as the weighted sum of the similarity between every two of their dominant feature vectors:

$$S = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{i,j} s_{i,j} \tag{6}$$

where the weighting factor $w_{i,j}$ is determined by the corresponding eigen-values:

$$w_{i,j} = \frac{\lambda_i}{\sqrt{\sum_{i=1}^{m_1} \lambda_i^2}} \frac{\sigma_j}{\sqrt{\sum_{j=1}^{m_2} \sigma_j^2}} \tag{7}$$

The weighting factor is such chosen for the following two considerations: 1) it should be proportional to the contributions of the corresponding dominant

feature vectors q_i and p_j ; and 2) This weighted sum should be equal to one, when two audio clips are the same, i.e., $q_i = p_j$ and $\lambda_i = \sigma_j$.

Actually, the dominant feature vectors are obtained from the covariance matrix of the frame based feature vector, and construct the base of the signal subspace. From this point of view, it can be considered that the similarity between two clips is in essence measured based on their noise-reduced covariance matrices.

3.2 Properties of the Similarity Measure

As mentioned above, the similarity between two clips is measured based on their signal subspaces, which can be obtained from the covariance matrices. Therefore, their similarity is actually a function of C_1 and C_2 , denoted as $S(C_1, C_2)$, where C_1 and C_2 are two covariance matrices of two clips, respectively. The properties of this similarity measure are discussed in this section.

Firstly, the similarity is symmetric when comparing two covariance matrices:

$$S(C_1, C_2) = S(C_2, C_1) \quad (8)$$

It is a basic requirement for most of similarity measure or distance measure.

Secondly, the similarity measure is normalized to the range from 0 to 1:

$$0 \leq S(C_1, C_2) \leq 1 \quad (9)$$

A larger value indicates more similar between the two clips. If two chips have the same dominant feature vectors and the same corresponding eigen-values, their similarity will be 1. Otherwise, if their dominant feature vectors are totally different, that is, orthogonal with each other, the similarity will be 0.

Moreover, the proposed similarity measure is robust. For example, if an audio clip is contaminated by superimposed or sequentially concatenated noise, its statistical features including mean and standard deviation will be affected much. However, its salient characteristics, or the corresponding dominant feature vectors, will not have significant difference. Therefore, our proposed similarity measure is not sensitive to the effect of noise.

For a better understanding of the proposed similarity measure, we explain it in a more intuitive way. As well known, one covariance matrix C determines a corresponding hyper-ellipse:

$$\{x : x^T C^{-1} x = 1\} \quad (10)$$

Assuming that q_i is one of the eigen-vector and λ_i is the corresponding eigen-value, thus q_i determines the orientation of one semi-axis of the hyper-ellipse, $\sqrt{\lambda_i}$ measures the length of the corresponding axis. The shape and orientation of the hyper-ellipse is an intuitive description of the characteristics of an audio clip. Fig. 2 illustrates two different hyper-ellipses in the 2-dimensional plane as an example.

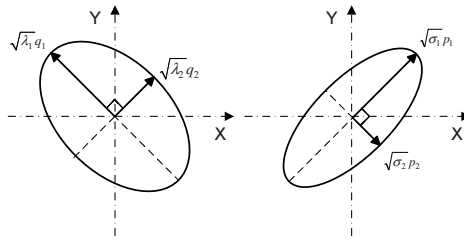


Fig. 2. Illustrations of two 2-D hyper-ellipses with two dominant feature vectors, a “fat” one is on the left, and a “slender” one is on the right.

In this way, the similarity between two audio clips can also be treated as the similarity of the shape and orientation between two corresponding hyper-ellipses. Thus, our proposed similarity measure can be deduced consequently, since the similarity between the orientations is measured by the inner-product of different dominant feature vectors, and the similarity between the shapes is affected by the weighting factor defined in Eq. (7).

3.3 Comparison with Other Distance Measures

In previous researches, several distance measures have been proposed and studied based on two covariance matrix [4], such as Kullback-Leibler in Eq. (11) and the Bhattacharyya in Eq. (12).

$$d_{KL}(C_1, C_2) = \frac{1}{2}(\bar{\mu}_2 - \bar{\mu}_1)^T(C_2^{-1} - C_1^{-1})(\bar{\mu}_2 - \bar{\mu}_1) + \frac{1}{2}tr(C_1^{-1}C_2 + C_2^{-1}C_1 - 2I) \tag{11}$$

$$d_{BHA}(C_1, C_2) = \frac{1}{4}(\bar{\mu}_2 - \bar{\mu}_1)^T(C_2^{-1} - C_1^{-1})(\bar{\mu}_2 - \bar{\mu}_1) + \frac{1}{2} \log \frac{\|C_1 + C_1\|}{2\sqrt{\|C_2C_1\|}} \tag{12}$$

where $\bar{\mu}$ is the mean of the sample vectors, and C_1 and C_2 are two covariance matrixes.

However, these distances utilize the inverse of covariance, and are usually used for the similarity measure between two sets of data, where a covariance matrix can be accurately estimated from sufficient data and represents the feature distribution of corresponding data set. In general, these distances are not suitable in measuring the distance between two audio clips, because of the following problems:

1. The covariance estimated from an audio clip is easily affected by noise.
2. If the duration of an audio clip is sometimes short and does not have enough sample frames, the estimated covariance matrix may not be full rank or ill conditioned. This will leads to numerical instability.
3. The noise-reduced covariance matrix can not be used directly in Eq. (11) and Eq. (12), since it is usually not full rank and thus not invertible.

The proposed similarity measure does not have these problems. Even if the original covariance matrix is not full rank, we can still extract the dominant feature vectors.

4 Experiments

In order to demonstrate the effectiveness of the proposed similarity measure, we compared its performance with some other similarity or distance measure, including L_2 distance, Kullback-Leibler distance and Bhattacharyya distance, based on a content-based audio retrieval system.

Our testing database consists of around 1000 audio clips. These sounds vary in duration from less than one second to about 30 seconds; and include many kinds of sounds, such as *animals, machines, vehicles, human, weapons* and so on.

In our experiment, all audio streams are down-sampled into 8 KHz, 16-bit and mono-channel, for universal processing. Each frame is of 200 samples (25ms), with 50% overlapping. Two types of features are computed for each frame: (i) perceptual features and (ii) 8 order Mel-frequency Cepstral Coefficients (MFCC). The perceptual features are composed of short time energy, zero crossing rate, pitch, 8 order subband energies, brightness and bandwidth. These features are then combined as a 21-dimensional feature vector for a frame.

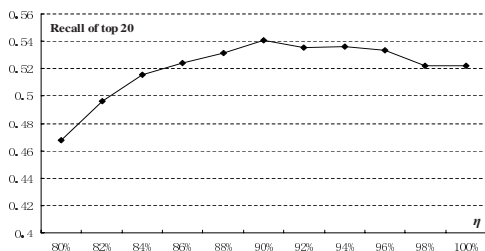


Fig. 3. The performance of the proposed similarity measure when the number of dominant feature vectors increase.

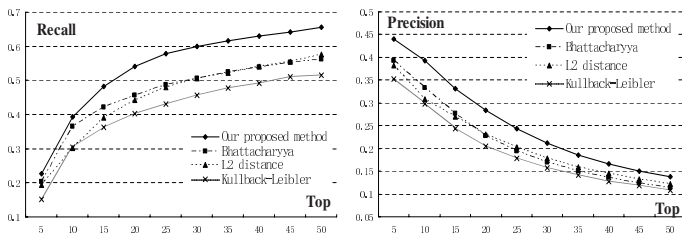


Fig. 4. Comparisons among the proposed similarity measure, L_2 distance, Kullback-Leibler and Bhattacharyya distance.

The first experiment is to find the best threshold η defined in Eq. (3), in order to decide how many dominant feature vectors should be used in the similarity measure. Fig. 3 illustrates a representative curve of recall ratio at the top 20 selection, indicating the influence of the threshold η from 80% to 100%. It can be seen that with the threshold increases, performance improves at first. The performance almost stops improving or even decreases when the threshold is more than 90%. The reason is that the first several dominant feature vectors contain important information of the clip, while the remaining is formed by noise effect and degrades the performance. In the following experiment, η is set as 90% to select the number of dominant feature vectors.

Fig. 4 illustrates the comparison results among the proposed similarity measure, L_2 , Kullback-Leibler and Bhattacharyya distance. In the experiments, L_2 distance is based on the mean and standard deviation of frame features of a clip; Kullback-Leibler and Bhattacharyya distance are both based on the covariance matrix. From Fig. 4, it can be seen that the proposed method has an obvious improvement, compared with the other similarity measures. For example, in the results of top 20, about 55% targets are retrieved with the proposed method, while only 45% is obtained using common L_2 distance. The corresponding results using Kullback-Leibler and Bhattacharyya are about 46% and 40% respectively. The precision is also increased compared with conventional similarity measures. In the results of top 20, the precision of the proposed method is about 28%, while other measure methods are less than 23%. The improvement is about 22%.

5 Conclusion

In this paper, a new similarity measure between audio clips is proposed. This similarity measure is based on the dominant feature vectors extracted from an audio clip. Compared with conventional mean and standard deviation, the dominant feature vectors give a better representation of an audio clip, especially when the characteristics change with time. Experimental results demonstrate the effectiveness of the proposed similarity measure. It also indicates that our approach is better than conventional L_2 distance, Kullback-Leibler distance and Bhattacharyya distance, in the case of similarity measure of two audio clips.

References

1. L. Lu, H.-J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation", *IEEE Trans. on Speech and Audio Processing*, 10(7):504-516, 2002.
2. E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based Classification, Search, and Retrieval of Audio", *IEEE Multimedia*, 3(3):27-36, 1996.
3. R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Using Structure Patterns of Temporal and Spectral Feature in Audio Similarity Measure", *Proc. of 11th ACM Multimedia*, pp. 219-222, 2003.
4. M. Basseville, "Distance Measure for Signal Processing and Pattern Recognition", *Signal Processing*, 18(4):349-369, 1989.

Generative Grammar of Elemental Concepts

Jose A. Lay¹ and Ling Guan²

¹ University of Sydney, Electrical and Information Engineering,
Sydney, NSW 2006, Australia,

² Ryerson University, Toronto, Canada
jlay@ee.usyd.edu.au, lguan@ee.ryerson.ca

Abstract. This paper presents a methodology on the use of elemental concept indices for operating semantic retrieval. In the methodology, the “language” by which semantics are communicated in audiovisual documents is explicated into a lexicon of elemental concepts and their generative grammar. Documents are then indexed with elemental concept indices such that more extensive querybility can be supported by means of post-indexing coordination of the elemental concept indices with the generative grammar operators.

1 Introduction

The defining characteristic of content-based retrieval (CBR) lies with the use of perceptual feature indices. In [1], nearly two dozens of perceptual features were identified to constitute the indices of some forty better known CBR prototypes reported in recent time. Thus, the state-of-the-art in CBR is on bridging the semantic gap. Perhaps a widely used approach is the hierarchical representation technique [2] where the search for “Lassy” a pet dog would be preceded by spotting the class “dog” that in turn comprised certain geometrical objects with certain patterns of colors, textures and shapes which are to be derived from perceptual features. Ideally the use of perceptual features would support arbitrary semantics. In reality, operating semantic retrieval directly upon perceptual feature indices has seen limitations on index exhaustivity. For example if *finger print* or *mammogram* documents are to be queried for content concepts; color, texture, shape, and layout features that work for general photographs will likely be unbecoming. Likewise, to search for similar faces, a feature like *eigenface* will need to be used. Thus operating a decent semantic retrieval system will require a multitude of perceptual feature indices. Furthermore, semantic information encompasses concrete and abstract objects. Attempts to resolve the semantic gap necessarily assume a consistent correlation between concepts and perceptual features. Sensibly, there are abstract concepts such as *cost*, *information*, and *time* that may not have perceptual representation. These concepts can not be supported naturally by perceptual features. Consequently, the current works have also seen limitation in querybility. They have been able to support only a small set of semantics on a rather small set of documents.

This paper presents a methodology for operating semantic retrieval of images and audiovisual documents by the use of elemental concept (elecept) indices. The conception of elecept and the working of this methodology are presented in Section 2. An exemplification on the retrieval of tennis game videos is presented in Section 3.

2 The Use of Elemental Concept Indices

In the generative grammar of elemental concepts methodology (G2E), a document is explicitly treated as a piece of information bearing medium having on it the representation of thoughts expressed in certain *languages*. Treating documents as the embodiment medium for expressed thoughts of certain languages provides an insight into how semantic retrieval may be best operated. As thoughts are communicated with a certain concept language in a document, the key to semantic retrieval is to allocate the lexicon and grammar of that concept language and to built index and query structures upon them. Perhaps one may draw a convenient analogy to the working of full-text indexing. In the parlance of the latter, semantic retrieval shall seek to allocate all distinct words of a language and the mechanism by which a *more specific concept* is expressed by using words in that language. Subsequently the words are used to index the document while the phrase building operators such as *concurrency* (AND) and *adjacency* (ADJ) are used to operate the query operation. Arbitrary concept queries may then be supported by post-indexing coordination of words with the phrase building operators. In the vernacular of [3], the basic design of G2E is to derive elecepts and generative grammar of the concept language and to build index and query structures upon them.

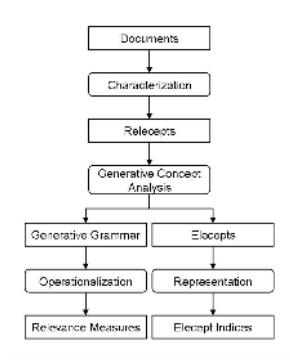


Fig. 1. Basic processes in G2E.

Fig. 1 illustrates the basic processes in G2E. First the salient aspects of relevance called *relecepts* of a database are identified. Each *relecept* is then subject

to a generative concept analysis where elecepts and generative grammar of the concept language are derived. Documents are then indexed with elecept indices, whereas generative grammar is used to operate the query operation.

The use of elecept indices extends several advantages. They serve as economical and consistent indices. As elecepts are finite discrete entities, a concise and uniform description scheme can be devised. For example, to account for semantic search of chess game, chess elecepts can be represented and indexed by using a shorthand notation. And unlike perceptual features which are medium dependent, the chess notation may be derived from a chess dictionary by means of character recognition, from a digital archive where a transformation operation is sufficient, or else from a video where more demanding vision-based mechanisms are required. Furthermore, they also extend less-demanding requirements on the operation for concept detection. As elecepts are granular concepts, the operations needed to detect elecepts are likely to comprise less demanding tasks. Sensibly to support the chess game concept, only the position of the chess pieces after each move is necessary; thus we need to process only those frames in a short moment after a move has taken place. Concept detection thus can enjoy the support of context hints to avoid the need to track the video continuously and expensively.

In operating the search, elecept indices can be treated as a vector of elecept terms where numerous techniques developed in modern information retrieval and CBR can be utilized. Meanwhile, unlike perceptual features, an inverted index of elecepts can also be judiciously devised. As elecepts are vocabulary of a concept language, indices of G2E comprise representation of granular concepts rather than simply perceptual features. For instance, the inverted index of elecepts can be used to find documents where certain movement concepts occur in chess games. The use of inverted index of elecepts in turn confers an attractive inducement, not only in terms of operational efficiency, but also the extension of queryability as various grammar operators such as AND, OR, NOT, ADJ, NEAR, WITHIN may be used to post-coordinate a large number of semantic queries from the fine granularity elecept indices.

3 An Exemplification on Tennis Video Retrieval

Several interesting works dealing with the retrieval of tennis games by using *keyword* and *perceptual feature* indices have been reported in recent years. In [4], an attempt was made to automatically generate high-level semantic annotations for tennis game. Concepts such as *baseline-rallies*, *passing-shots*, *net-games*, and *serve-and-volley games* were detected and indexed as keywords to facilitate semantic querying. Example queries envisioned in the work were: (1) to retrieve *serve-and-volley* clips that contain *John McEnroe* in 1984; (2) to retrieve all *baseline-rally* clips that contain *Andre Agassi* on a hard court in 1992. In [5], the work was extended to augment the indices with action concepts such as *forehand stroke*, *backhand volley*, and *smashing* by also detecting the ball positions and certain behaviors of the players. Semantic annotations are represented and

indexed as keywords. In [6], an attempt was made to recapture the querying flexibility of the post-indexing coordination scheme by directly deriving semantics from perceptual feature indices. In the work, document was indexed by using a set of perceptual features consisting of the player positions; dominant colors; and shape features such as mass center, area, bounding box, orientation, and eccentricity. The retrieval task was then operated as a combination of operations based on the rule-based grammar and the stochastic processes. The rule-based grammar was used to derive the spatial events and objects such as the *player_near_the_net* event; while the stochastic process was used to detect temporal events such as *service*, *smash*, *forehand volley*, and *backhand volley*.

Clearly the use of annotation indices allows concept search to be operated as *cataloging* or *keyword* search but is unable to provide high exhaustiveness and queryability and lacks support for fine granularity queries. Supported queries are limited to coordination of indexed semantic labels. On the other hand, to support semantic retrieval by search time processing of low-level perceptual features poses a very high computation cost. The latter appeared as the limiting factor when support for a larger set of game concepts over a larger database is desirable.

3.1 Tennis Concept Language

A multitude of similarity aspects exists in tennis game. One way to articulate the significant relecepts is to take into account numerous usage contexts. Those aspects are a matter of different importance to different attentions. A commentator holds interest primarily on game statistics; a viewer may prefer to watch only game highlights; and a coach requires access to skills and strategies while a fashion designer may search for all about figure-hugging tennis wear. For the purpose of this exemplification, we focus only on the sports concepts and categorize them into four classes: *competition concepts*, *action and skill concepts*, *game strategy concepts*, and *game statistics*.

To deal with those game concepts, we assume the existence of a *tennis concept language* (TCL) by which the game concepts are communicated. More specifically TCL comprises a finite tennis elecept lexicon along with its generative grammar. A tennis game concept thus can be seen as a phrase or sentence constructed out of a finite series of tennis elecepts coordinated by rules of the tennis concept grammar. The operation by G2E is thus to identify, represent, and index a tennis video with these tennis elecepts and to embed the rules of the grammar into the query operation.

To derive the tennis elecepts, we structure the numerous tennis game concepts along the competition hierarchy in a way similar to [7]. A tennis match is organized as a hierarchy of match-set-game-point and stroke where salient concepts at each level of the hierarchy are established. At the levels of match, set, game, and point; the salient concepts are associated with the information about the competition scores whereas at the stroke level, the salient concepts are those associated with the trace of the ball and the action skill of the player: *where the ball was hit*, *where it was returned*, *where the players were*, and *how the ball was hit*. The first three are salient concepts for characterizing the game strategy

concepts while the last one deals with the action concepts. Three sub languages of TCL can thus be substantiated. The competition concepts are expressions of the score language. The strategy concepts are expressions of the position language while the action concepts are expressions of the player's posture language. This exemplification will further focus on the conception and organization of the position language upon which retrieval for many exciting skill and strategy concepts is facilitated. Having explicated the niche scope, we refer to the position elecepts simply as the tennis elecepts.

3.2 Tennis Game Tiles Notation

In the earlier works, movement of ball and players has been maintained as continuous motion trajectories. These trajectory features result in rather complex index structures that are not so practical for retrieval purposes. On the other hand it is understood that some segments within the movement trajectories are more important than others with respect to the characterization of the tennis game concepts. The tennis elecepts identified in the preceding section comprise only information about where the ball was hit, where the other player stood, and where the ball was returned. The tennis elecepts are discrete conceptual structures that will be best presented as is.

To represent these tennis elecepts, we segment the tennis court into a board of tiles. The tiles can be segmented in various forms to suit certain dissimilar purposes. The simple form used in this work is illustrated in Fig 2(a). Once the game space is segmented into tiles, movement of ball and players can be expressed by using a *game tile notation*. For example, the movement of ball in a particular point can be recorded by the sequence of its landing position: D1-C6-B1-E8. In this example, the ball was first served from D1 to the far right bracket of the service court C6. It was then returned straight to the far left end of the baseline B1, and then terminated by a zigzag return to E8. Other tennis elecepts can be recorded in similar manner. To represent the tennis elecepts, a triplet (l,p,h) comprising the ball landing spot (l), the ball hit location (h), and the other player position (p) at the time the hit occurred is used in this work. In the latter the notation (*, D1, C9)-(C6, B8, C2) denotes that the ball was served from D1, at the time the opponent player was at C9. The ball then landed at C6 then bounced towards and was returned from B8, at that time the serving player was at C2. In this tile notation, the umpire stand has been used as the reference point where the upper left corner tile faced by the umpire is set to tile A0.

By using the game tile notation numerous tennis elecept indices can be represented discretely. Furthermore subject to the detection setup one can also add accuracy by also registering information about the height of the ball when it was hit, the time, and the posture of the players of importance to the action skill concepts. Certainly the latter will also requires the posture language to be developed.

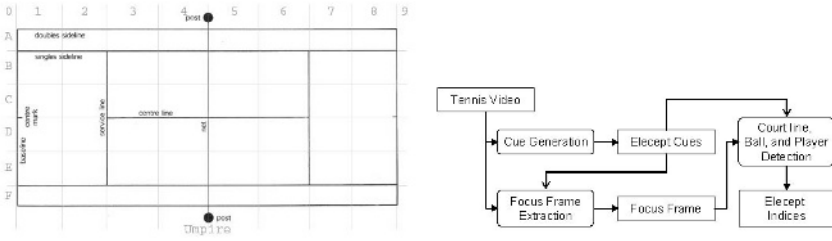


Fig. 2. *left* (a) Game space for movement notation of the game of singles. *right* (b) Detecting elecept indices by using context hints.

3.3 Tennis Elecept Indices

To circumvent the complexity and inflexibility respectively associated with the motion trajectory features and the semantic annotation indices, a tennis video is indexed by using tennis elecept indices in G2E. In the basic form, a tennis video can be surrogated by competition scores along with the tennis elecepts represented by using the game tile notation. An illustration of the elecept indices for the seventh game of the 2002 US Open final featuring Agassi and Sampras is shown in Fig 3.

The elecept indices are beneficial to the operation of semantic detection. As tennis elecepts are simple granular concepts, elecept indices can be derived by using more manageable detection operations. To detect the tennis elecepts in tennis videos, audiovisual *context hints* can be used. As we are interested only in the position information at the time when the ball touches ground or being hit by the player, the elecept detection may center only on frames where those events occurred. Intuitively, the “tag” and “tog” sounds associated with the *ball-hit* event and the *ball-land* event in a typical tennis video can be used as elecept cues to serve as potent indicators for identifying the occurrence of the *ball-hit* and *ball-land* events. Additionally audio cues associated with the *net*, *fault*, and *out* events along with visual cues identifying the numerous *wide-angle court* views and the side of the umpire stand are used. The tennis elecept detection procedure is illustrated in Fig 2(b).

In addition, elecept cues can also be used in groups for numerous purposes. First in the absence of a confident cue, such as the failure to detect a ball-land cue where two consecutive ball-hit cues were detected, the two ball-hit cues can be used as the space limiter for constraining the visual processing of the ball-land detection to the proper group of frames alone. The visual detection may then lead to the ball-land event being detected or else the ball was returned without ever touching the ground. Then they can be used in association with domain knowledge. For example, it is safe to assume that a ball-land event be preceded by a ball-hit event. Thus, if only a ball-land event was detected, one can assume that the preceding ball-hit event had been omitted or missing in the segment. Similarly cases can be made on the service-fault event and the ball-net event.

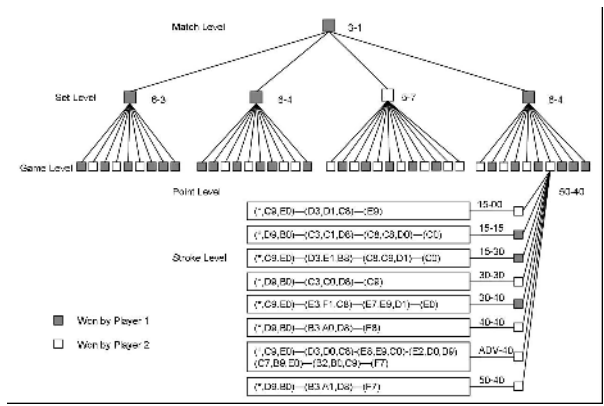


Fig. 3. Elecept representation of a tennis match.

Once focus frames are extracted and their elecept contexts established, elecept indices can be derived by using the techniques developed in [4][5].

3.4 Inverted Index of Elecept Indices

An inverted index can be constructed for each of the elecepts—ball-hit, ball-land, and player-position. An inverted index $E^{-1}=(e_i, s_j, p_k, t_n)$ is defined here by a tuple composed of *vocabulary* and *occurrences*. The vocabulary comprises the elecept terms e_i , while the occurrences comprise the identity of the point segments s_j , the relative sequence position p_k , and the time t_n when the elecept term appears in the score point. The inverted indices of ball-hit elecept for the point segments in Table 1 is illustrated in Table 2.

Table 1. Example point segments.

Segment	ID	Portable Description
[00-00]	1	(*C9E0) (D3D1C8) (E9)
[15-00]	2	(*D9B0) (C3C1D8) (C8C8D0) (C0)
[15-15]	3	(*C9E0) (D3E1B8) (C8C9D1) (C0)
[15-30]	4	(*D9B0) (C3C0D8) (C9)
[30-30]	5	(*C9E0) (E3F1C8) (E7E9D1) (E0)
[30-40]	6	(*D9B0) (B3A0D8) (F8)
[40-40]	7	(*C9E0) (D3D0C8) (E8E9C0) (E2D0D9) (C7B9E0) (B2B0C9) (F7)
[AD-40]	8	(*D9B0)(B3A1D8) (F7)

The inverted index of elecepts can be used to support a number of content concept queries. In the simplest form, it can be used to collocate for individual elecept terms. For example, one may search for game segments where the ball landed on the lower right corner of the court. A look up for E8 on the inverted index of ball-land elecepts results in point segment 7 being selected. Similarly one may search for where the ball was hit or where the player was standing.

Table 2. Inverted indices for the ball-hit elecept.

No	Ball-Hit	(Segment: Pos: Time)
1	A0	(6;2;110)
2	A1	(8;2;101)
3	B0	(7;6;504)
4	B9	(7;5;354)
5	C0	(4;2;105)
6	C1	(2;2;108)
7	C8	(2;3;159)
8	C9	(1;1;000) (3;1;000) (5;1;000)(7;1;000) (3;3;203)
9	D0	(7;2;059) (7;4;301)
10	D1	(1;2;103)
11	D9	(2;1;000) (4;1;000) (6;1;000)(8;1;000)
12	E1	(3;2;058)
13	E9	(5;3;158) (7;3;209)
14	F1	(5;2;103)

Certainly the collocation can also be coordinated to form structured concept queries. Let us represent ball-hit with X, ball-land with L, and player-position with P. The user may want to search for game segments where the ball landed on any corner of the court. In this case the search is operated by LB1 OR LB8 OR LE1 OR LE8 on the inverted ball-land indices. By extension a structured concept query can be specified for numerous inverted indices. Instead of searching only the ball-land position, for example, the user may search for game segments where the ball was hit from D0 at a time when the opposite player was not standing right behind the service line and the ball then landed on E8. Sensibly the query can be expressed as XD0 AND NOT (PB7 OR PC7 OR PD7 OR PE7)) AND LE8.

As the position of occurrence of elecept terms is also indexed, proximity concept queries can also be supported. Two proximity operators are widely used in text retrieval. The adjacent (ADJ) operator searches for adjacent words such as “Information Retrieval” while the within (W) operator is used to search for two particular terms that occur within a specified number of words in the text. The ADJ and W operators can also be supported in the inverted index of elecepts. When posing the previous query, we assume that the ball that landed on E8 was indeed the one that was hit from D0. This relationship can be more stringently enforced by using the ADJ operator, the query thus become XD0 ADJ LE8. The ADJ operator is operated by using the “Segment” and “Pos” information in the inverted indices. More specifically, two elecept terms are said to be in adjacent relationship if they occurred in consecutive sequence in a score point. Furthermore, the query can be further restricted by ascertaining that the term XDO needs to take place before the term LE8 i.e. that Pos XD0 ; Pos LE8. Further still, more extensive querybility can be supported by also introducing the W operator. In this case, a user may also search for the pattern of game-play where the player drove a crosscourt corner ball and within three strokes delivered a drop shot. Their representation will allow even more concept queries to be operated meaningfully and efficiently.

As elecept indices are discrete semantic units comparable to words in a text document, the inverted file of tennis elecept indices can be substantiated practically allowing keyword (elecept) search to be facilitated by using techniques

traditionally developed for keyword search. Furthermore, as detection is carried out only on the indexing stage, the costly pattern recognition task is contained to the indexing process. On the other hand, in comparison with the semantic annotation indices, the elecept indices can support higher specificity query and allow for greater flexibility. Instead of being confined to coordination of a few semantic labels, the elecept indices allow a rich set of granular concept queries to be operated.

4 Conclusions

In this paper, we presented the use of elecept indices for semantic retrieval. We are convinced that once a significant aspect of similarity is determined, the post-indexing coordination of elecept indices in G2E can render accessible more extensive queryability than systems based on perceptual feature indices.

References

1. R.C. Veltkamp, M. Tanase, and D. Sent: *Features in Content-based Image Retrieval Systems: A Survey*, A chapter in State-of-the-art in Content-Based Image and Video Retrieval, edited by R.C. Veltkamp, H. Burkhardt and H.-P. Kriegel, Kluwer, 2001, pp. 97-124.
2. R. Mehrotra: *Content-based Image Modeling and Retrieval*, Proceedings of the Clinic on Library Applications of Data Processing, University of Illinois at Urbana-Champaign, March 1996, pp. 57-67.
3. J.A. Lay and L. Guan: *Retrieval for Color Artistry Concepts*, IEEE Transactions on Image Processing, 13 (3), March 2004, pp.
4. G. Sudhir, J.C.M. Lee, and A.K. Jain: *Automatic Classification of Tennis Video for High-level Content-based Retrieval*, Proceedings of the IEEE International Workshop on Content-based Access of Image and Video Database, 1998, pp. 81-90.
5. H. Miyamori and S.-I. Iisaku: *Video Annotation for Content-based Retrieval using Human Behavior Analysis and Domain Knowledge*, Proceedings of the fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 320-325.
6. M. Petkovic and W. Jonker: *Content-based Video Retrieval by Integrating Spatio-temporal and Stochastic Recognition of Events*, Proceedings of IEEE Workshop on Detection and Recognition of Events in Video, 2001, pp. 75-82.
7. L. Jin and D.C. Banks: *Visualizing a Tennis Match*, Proceedings of the IEEE Symposium on Information Visualization '96, 1996, pp. 108 -114, 132.

Learning Image Manifold Using Web Data*

Xin-Jing Wang^{1,2}, Wei-Ying Ma¹, and Xing Li²

¹ Microsoft Research Asia, Beijing 100080, P.R. China
wyma@microsoft.com

² Department of Electronic Engineering, Tsinghua University,
Beijing 100084, P.R. China
wxj01@mails.tsinghua.edu.cn
xing@cernet.edu.cn

Abstract. Manifold learning has become a hot research topic in recent years and is widely used in the area of dimension reduction, information retrieval and ranking, etc. However, how to reconstruct the intrinsic manifold from the observed data points, i.e. what is the proper data point distance measure, is still an open problem. In this paper, we propose to take advantages of the information provided by web-pages and the image-related website link structure to learn the Web image manifold, which better approaches to the intrinsic manifold than those learned by previous methods which use Euclidean alike distances to construct the initial affinity matrix. Experimental results prove the effectiveness of our learned Web image manifold.

1 Introduction

The rapid growth of data has brought many troubles to both storage and processing for many applications, such as computer vision, information retrieval and text mining, since these data are typically of high-dimension. However, in many cases, these data points can be considered as lying in or close to a low-dimensional embedding of the high-dimensional space. Learning the nonlinear low-dimensional structures hidden in a set of unorganized high-dimensional data points is known as manifold learning [1,5,7,8,11,13].

A manifold is a topological space which is locally Euclidean. On the manifold, the distance between two data point is measured by the geodesic distance (the length of real line in Fig.1) rather than the Euclidean distance (the length of dotted line in Fig.1). Hence, the intrinsic manifold is human concept-based and without semantic gap. However, how to exploit the geodesic distances and precisely simulate the underlying nonlinear manifold is still an open topic. [11] proposes a global approach called Isomap to reconstruct the underlying manifold. For each pair of neighboring data points, it finds the shortest path under the constraint that the path must hop from neighbor to neighbor. And the geodesic distances are approximated by the length of paths. It seeks to map nearby points on the manifold to nearby points in low-dimensional space, and faraway points

* This work is done when Xin-Jing Wang is an intern in Microsoft Research Asia.

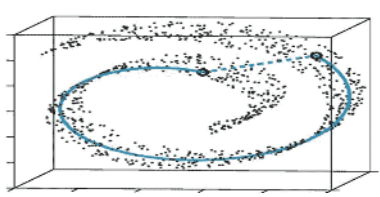


Fig. 1. An Example of Data Manifold: the “Swiss-roll”. Dots represent the data.

on the manifold mapped to faraway points in low-dimensional space [8]. [1,7] are local approaches which essentially seek to map nearby points on the manifold to nearby points in low-dimensional space. Such kind of approaches have good computational efficiency and representational capacity [8]. However, there exists an underlying assumption for these approaches, i.e. the data points are dense enough to ensure the approximation of manifold structure. When the number of sample points is too small to describe the underlying topology of the data, the geodesic distance on the manifold may not be accurately estimated [5]. In order to solve this problem, [5] proposes to make use of user’s relevance feedback to modify the similarity measure in image retrieval: either shorten the distances between the query image and positive images, or lengthen that between the query and negative images.

In this paper, we propose to leverage the Web data as an effective way to recover the intrinsic Web image manifold which performs much better than the previous approaches using content-based similarity measures.

The manifold is learned in two steps: Firstly, we automatically identify the concepts of Web images based on the analysis of corresponding web-pages. Then the image links are analyzed to semantically organize these Web images. The geodesic distances between each pair of images are thus defined as the shortest path on the learned image concept hierarchy, on which the intrinsic Web image manifold is effectively approximated.

The paper is organized as follows: in Section 2, we propose the image auto-annotation approach which obtains the initial set of image concepts. In Section 3, the website link structure extraction approach is discussed, based on which the images are organized. Section 4 discusses the manifold learning approach. We give our experimental results in Section 5 and conclude this paper in Section 6 with a discussion of future works.

2 Associating Image Contents with Concepts

As Web images are typically surrounded by abundant textual annotations, they can be considered as labeled data. Assume each image consists of two parts: a key region representing the main concept and environmental regions representing the context, we intend to extract the right keywords and associate them with the corresponding key regions in the images. The selected keywords are called

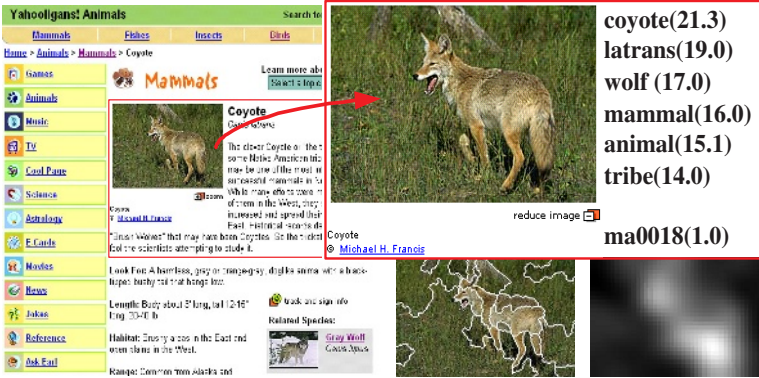


Fig. 2. Term extracted for a Web image and its saliency map.

key terms hereafter. The resulted (key term, key region) associations serve as a vehicle to map image content into human concepts.

The Web images are automatically annotated in a two-step way: 1) key term extraction 2) key region extraction. Fig.2 shows an example of a web-page (the left part). Based on a vision-based web-page segmentation approach [2], we can obtain a cleaner set of surrounding texts for the Web image (see the red rectangle in the left part of Fig.2). Then HTML tags and some heuristic rules are used to weight these texts through which the top-ranked keyword is selected as the key term [12], as shown in the top-right part of Fig.2.

On the other hand, each image is segmented and passed through an Attention Model [6] for the key region extraction. We order the regions based on their attention values. An example of image attention map is shown in the bottom-right of Fig.2. As can be seen, it helps us identify the “coyote” region from the “grass” background. The most salient region is selected as the key region to associate with the key term. By this way, we obtain a large collection of key regions and the associated key terms that are very likely to be the semantic annotation of these regions.

3 Organizing Image Concepts by Link Analysis

In order to both reduce the incorrect recognition and aggregate the similar concepts, we take advantages of website links to organize the learned image concepts hence the images. [12] makes use of WordNet hypernyms to structuralize the learned (key term, key region pairs). Although it appears to be an effective way, there are two drawbacks: 1) it will result in a large number of useless nodes in the tree which increase the complexity of geodesic distance calculation; 2) only the information provided by the isolated web-pages is made use of. In fact, web-page hyperlinks contain abundant information such as topic locality and anchor description [3]. Topic Locality means the web-pages connected by the same

hyperlink are more likely of the same topic than the unconnected web-pages. And anchor description means that the anchor text of a hyperlink is often a good summary of the target web-page. Because a website designer usually conceives the website information structure before s/he constructs it, if all target web-pages are replaced by their corresponding anchor texts, these anchor texts are then topic-related. Hence it is possible to obtain a semantic network from the Web's link structure, in which image concepts are the leaf-nodes, anchor texts are other immediate nodes and semantic relations are edges. Our goal is to extract such a latent image concepts structure from the website link structure, which in theory reflects human's view on concepts relationships.

Web Link Analysis has been applied in many applications. [10] analyzes the structural information encoded in URL to understand the website designer's intention. [3] separates semantic links from navigational links and applies link analysis to each website to obtain its content structure. In our approach, we first select a set of seed images as starting points to filter out noisy links. Then the anchor texts of hyperlinks for each website are extracted as immediate nodes and their relations as edges to form semantic trees, one tree for each website. These trees are merged using WordNet [4] which finally results in organized image concepts. The leaf nodes are the Web images as the data points to reconstruct the underlying manifold.

3.1 Starting Point Selection

Based on all the training Web images which have their key terms successfully identified, we perform some simple while effective heuristic rules to select a subset of images as the starting point to extract their corresponding website content structure. An image is not a seed image if

- The image is not placed in the central block of a web-page. We assume that an image is center-positioned if the block (i.e. web-page segment) contains it has a high importance value [9].
- The aspect ratio of the image exceeds 2:1 (or 1:2). Usually such images tend to be advertisements or navigational logos, etc.
- The image has no uplinks but has down-links. Such images are usually thumbnails or entry points which have the same function as anchor texts.

The images left form the seed image set whose uplinks are analyzed for the website image concept structure extraction.

3.2 Image-Related Website Link Structure Extraction

Based on the seed images obtained in Section 3.1, we use a backward tracing method to obtain the entire website structure related to the corresponding Web images. Firstly, we group the seed images according to their URLs. Each group of images corresponds to a certain website. Starting from each group of images, we extract their related website structure by filtering the noisy navigational uplinks and keeping the semantic uplinks according to the following heuristic rules:

- if a hyperlink uplinks to an image, save its surrounding texts, reserve this link and keep tracing back. This is because such target images normally serve as either entries to a set of images with similar concepts or as thumbnails;
- if a hyperlink uplinks to an anchor text, check if this anchor text is noun. If so, save the anchor text, reserve this link and keep tracing back; else remove this image.

This approach is performed iteratively until the homepage is reached. In such a way, we obtain a semantic tree (or several trees) for each website whose leaf nodes are seed images.

3.3 Merging Website Trees

We merge the website trees to one single semantic tree. Let X and Y be two trees, $X_{i,j}$ denote the concept of j^{th} node in i^{th} level of tree and $Y_{l,k}$ denote the k^{th} node in l^{th} level of tree. For the leaf nodes, $i = 0$. The father of $X_{i,j}$ is denoted as $X_{i+1,j}$ and $X_{i-1,j}$ a child. Similar definitions hold for Y . Assume X is larger than Y and we attempt to merge Y to X , the bottom-up merging strategy is shown in Fig.3. Those isolated images can be seen as single-node tree and processed in the same way as Fig.3.

- 1) (Initialize) $i = 0, l = 0$
- 2) if $X_{i,j} = Y_{l,k}$ or $X_{i,j} \approx Y_{l,k}$, i.e. $Y_{l,k}$ is a synonym of $X_{i,j}$ by WordNet, merge $X_{i,j}$ and $Y_{l,k}$
- 3) if $X_{i,j}$ is a hypernym or its synonym of $Y_{l,k}$, insert $Y_{l,k}$ into the children set of $X_{i,j}$
- 4) if $Y_{l,k}$ is a hypernym or its synonym of $X_{i,j}$, set $X_{i,j} = X_{i+1,j}$ and go to step 2). After $Y_{l,k}$ is inserted into X , set back $X_{i,j}$
- 5) if all the siblings of $Y_{l,k}$ are processed, set $Y_{l,k} = Y_{l+1,k}$ and go to 2)

Fig. 3. The Pseudo-Code of Tree Merge Algorithm

Possibly there are several trees resulted after merge. We create a pseudo-root node named “entity” to combine these trees into a single one.

4 Discovering Intrinsic Image Manifold

As mentioned above, the website link structure reflects the designer’s view about the relationships among different image concepts. Because the images (concepts) are organized supervised by human-constructed website links, we can use the learned concept tree to discover the underlying Web image manifold which reflects the concept-based image relationships.

We adopt the approach in [13] to rank images given a query. In [13], the initial affinity matrix W is calculated based on Euclidean distance which may

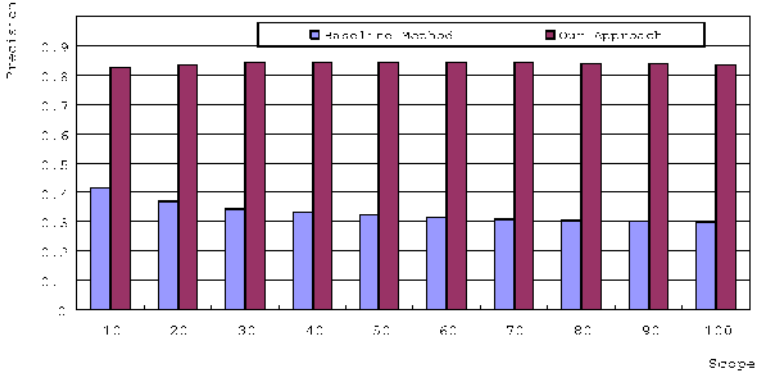


Fig. 4. Retrieval Precision Comparison.

degrade the learned manifold due to the semantic gap, we modify W according to the learned concept tree as below. Let W_{ij} be an element of W . In [13], $W_{i,j} = \exp[-d^2(x_i, x_j)/2\sigma^2]$ where $d^2(x_i, x_j)$ is the Euclidean distance between data x_i and x_j . We use the image concept tree to bridge the semantic gap in calculating $d(x_i, x_j)$.

Intuitively, the farther two nodes are in the tree, the less similar their concepts will be. Assume there are M leaf nodes and L nodes in the tree, we assign each edge of the tree a weight of $1/L$. The new distance $d^*(x_i, x_j)$ is given by

$$d^*(x_i, x_j) = \frac{k}{L} \times d(x_i, x_j), \quad 1 \leq i, j \leq M \quad (1)$$

where k is the shortest path between x_i and x_j according to the tree. Equation (1) means: If x_i and x_j are two Web images, use the short path on the tree (i.e. the concept-based distance) to weight their content-based similarity measured based on Euclidean distance as the estimation of the intrinsic geodesic distance. The nearer two images are on the tree, the less their distance is weighted.

In such a way, the pairwise image distances are modified to form the affinity matrix W and the algorithm presented in [13] is adopted to learn the manifold and support image retrieval or annotation given a query image.

5 Experiments

We crawled about 17,000 JPEG images from the Web and 5148 images are identified as seed images to extract the website link structures. We extract the 36-bin color correlograms as image region low-level features.

All the seed images are used as queries to evaluate the retrieval performance on the learned manifold. We use the method proposed in [13] as the baseline. The precision-scope performance measure is applied. Scope specifies the number of images returned to the user in each retrieval round. Precision is defined as the number of retrieved relevant objects over the value of scope.

Fig. 4 shows the precision comparison of our method and the baseline. The two parameters used in manifold learning [13] is $\sigma = 0.6, \alpha = 0.8$. It can be seen that significant improvement is achieved by our approach. It means that our proposed distance measure, i.e. modifying the content-based distance according to the concept relationships learned by web-page analysis and website link analysis, can effectively approximate the geodesic distance, hence recovers the intrinsic manifold structure of the image manifold.

Note that only the key region of each query is currently used, and environmental information is removed (in fact, the correlation information can also be used to improve the performance [12]).

6 Conclusion and Future Works

We propose an idea of using Web data, i.e. Web image textual annotations and website link structure, to approximate the underlying Web image manifold. The Web images and their textual annotations are used to learn an automatic map from image low-level features to high-level semantic concepts and the image-related website link structures are taken advantages of to organize the image concepts. Based on the learned image concepts structure, the traditional content-based image distance measure can be modified according to their semantic distance, which better approximate to the intrinsic geodesic distance on the underlying manifold. We show its effectiveness in image understanding.

Since our method is based on [13], a disadvantage of our current approach, which is also the disadvantage of many current manifold learning approaches [1,7,11,13], is that the manifold learned is defined only on the training data points. It is not discussed in this paper how to map the new test points into this manifold. We will research on this in our future works.

Acknowledgement. Sincerely acknowledgement should be given to Jian-Tao Sun for his help on preparing the dataset for experimental evaluation.

References

1. Belkin, M., Niyogi, P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, In T.G.Dietterich, S. Becker and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems* **14**. Cambridge, MA: MIT Press (2002)
2. Cai, D., Yu, S.P., Wen, J.R., Ma, W.-Y.: VIPS: a Vision-Based Page Segmentation Algorithm. Microsoft Technical Report, msr-tr-2003-79, (2003)
3. Chen, Z., Liu, S.P., Liu, W.Y., Pu, G.G., Ma, W.-Y.: Building a Web Thesaurus from Web Link Structure. *SIGIR* (2003).
4. Fellbaum, C.: *WordNet: An Electronical Lexical Database*. MIT Press, Cambridge, Mass. (1998)
5. He, X.F., Ma, W.Y., Zhang, H.J.: Learning an Image Manifold for Retrieval. *ACM Multimedia* (2004)

6. Ma, Y.F., Zhang, H.J.: Contrast-based Image Attention Analysis by Using Fuzzy Growing, *ACM Multimedia* (2003).
7. Roweis, S., Saul, L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science* **290**(2000), 2323-2326
8. Silva, V., Tenenbaum, J.B.: Global versus Local Methods in Nonlinear Dimensionality Reduction. *Neural Information Processing Systems* **15** (2003), 705-712
9. Song, R.H., Liu, H.F., Wen, J.R., Ma, W.Y: Learning Block Importance Models for Web Pages, *Proceeding of the Thirteenth World Wide Web conference* (2004), 203-211
10. Spertus, E.: ParaSite: Mining Structureal Information on the Web. In *Proc. of WWW6* (1997), 587-595
11. Tenenbaum, J., Silva, V., Langford, J.: A Global Geometric Framework for Non-linear Dimensionality Reduction, **290** *Science* (2000), 2319-2323
12. Wang, X.J., Ma, W.-Y., Li, X.: Data Driven Approach for Bridgin the Cognitive Gap in Image Retrieval. *IEEE Conf. on Multimedia and Expo* (2004)
13. Zhou, D.Y., Weston, J., Gretton, A., Bousquet, O., and Schölkopf, B.: Ranking on Data Manifolds. In: Thrun, S., Saul L. and Schölkopf, B. (eds.): *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, Mass

Novel Concept for Video Retrieval in Life Log Application

Datchakorn Tancharoen and Kiyoharu Aizawa

Department of Electrical Engineering, Graduated School of Engineering
Department of Frontier Informatics, Graduate School of Frontier Science
The University of Tokyo
707 Kiban-Building, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561 Japan
{dtan,aizawa}@hal.k.u-tokyo.ac.jp

Abstract. At present in daily life, many people prefer to record their experiences in multimedia forms instead of by writing a diary. Hence, we have developed the Life Log system to record and manipulate our experiences efficiently. Content-based features from audiovisual data are necessary to detect the significant scenes from our life. One important group of scenes is conversations that contain useful information. However, content-based features alone cannot satisfy people's preferences. This paper demonstrates a novel concept in video retrieval: integrating the content of video data with context from the various sensors in the Life Log system. We attempt to extract the important features from audiovisual data and context features from wearable devices to detect interesting conversations. The experiments present conversation scenes based on audiovisual features and the additional contexts to support more semantic conversation analysis.

1 Introduction

Recently, imaging devices such as the digital camera and video camcorder have developed rapidly. As a result, many people would like to record their life in the form of images and video. Therefore, such data have become important tools in daily life. However, the huge amount of data in visual information means that much time is consumed in searching for an interesting image in a person's life record. To capture our experiences, the Life Log system is being developed to record and manage our entire experiences efficiently [1].

Retrieval techniques are necessary to make it possible to effectively search for interesting scenes in one's life record. Content-based video retrieval is an active area of research that enables us to acquire many useful features and efficiently retrieve important information. There are some studies in this area [2], [3].

In the Life Log application, we want to retrieve interesting scenes and acquire important information from them. These significant scenes are not easily retrieved using general retrieval methods. In addition, existing content-based frameworks do not satisfy all the demands created by people's subjectivity. Hence, we must use both content-based features and other data from wearable

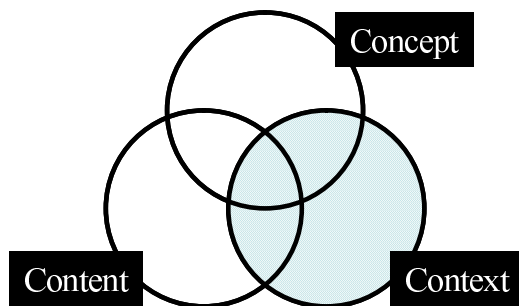


Fig. 1. The relationship of content, context, and concept-based methodology

devices that were recorded synchronously with the audiovisual data in the Life Log system.

This novel concept of video retrieval is an interesting topic. Its basis is the integration of frontier informatics technologies, and includes content-based processing, sensor technology, database systems and indexing algorithms. For example, we may wish to use a complex sentence to search for an interesting scene in our life such as “I would like to find the scene when I was talking with my friend in the sushi restaurant near my university during lunchtime”. It looks really fantastic, but our approach makes it possible.

Our viewpoint of video retrieval is sketched in Fig. 1. Current research in the content, context, and/or concept areas includes content-based processing, sensor technology, and human-computer understanding. In this work, we are concentrating on the context area, which includes some aspects of content and concept to support our novel retrieval technique. We are also working on audiovisual feature extraction from the content-based viewpoint. However, concept-based retrieval also includes such high-level concepts as face recognition and speech identification, which require more complex methodology and computational effort. Therefore, our current research concentrates on the context, content and low-level concept areas to support people’s preferences in the Life Log application.

This paper presents the novel concept of video retrieval for Life Log application based on the content of audiovisual information and context data from various sensors. Section 2 introduces our current Life Log system. The content-based retrieval technique for conversation scene detection is presented in Section 3. Section 4 describes the additional context information from Life Log data. The novel retrieval method and discussion are given in Section 5. Finally, Section 6 is the summary of this paper.

2 Life Log System

In the Life Log application, we try to record our entire life in the form of multimedia information and try to manipulate the data efficiently [1]. Multimedia information can be recorded by a wearable camera, a microphone, and wearable

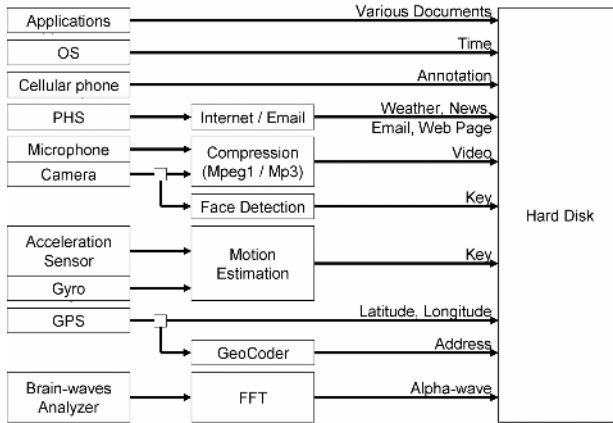


Fig. 2. Block diagram of various types of data in current Life Log system

devices including a brain wave analyzer, GPS receiver, motion sensors and information from a portable PC. The information from wearable devices is transferred to the PC and processed to support the Life Log application as shown in Fig. 2.

2.1 Novel Concept Direction

There are many advantages of content features that provide local information directly from multimedia data. Nevertheless, content features from audiovisual data alone cannot always match users’ preferences and the concept of the Life Log application. Hence, our research concentrates not only on content, but also on context as a novel concept-based retrieval method. Context can provide much information that cannot be obtained from content, including location, human feeling, and the user’s movement.

We are developing the Life Log application based on the integration of content and context features to increase system efficiency and to try to acquire interesting scenes based on the user’s preferences. The details of content-based retrieval using audiovisual data and data from various wearable devices are described in the following sections.

3 Content Based Retrieval Algorithm

Content-based information can be acquired from audio and visual data that are recorded from a microphone and wearable camera. The video data provide visual information and sound from the microphone will provide the audio data, which together comprise the contents of Life Log system. Conversation scenes can be important scenes that include many interesting topics. Thus, this section will describe conversation scene detection as an example of content-based retrieval in the Life Log application.

Existing techniques for human speech detection [4], [5] require training data that incorporate high-level audio features and pose a high computational load. However, we require a low-complexity method with high performance to detect interesting scenes. Hence, the low-level features of audio and video data will be analyzed for identifying conversation scenes [6].

In addition, we usually concentrate on a person's face with eye contact when we converse with other people. A wearable camera should therefore capture faces during conversations. From this visual information, face extraction based on skin color is executed [7], [8]. The techniques of conversation scene detection are explained in the following sections.

3.1 Human Voice Detection

A conversation scene contains audio data as the main detection key. However, audio data include other sounds as well. Hence, voice detection is based on characteristics of the human voice including power spectrum, fundamental frequency, and speech continuity.

Silent scenes are discriminated by power spectra. At the same time, unpredictable noise is removed by detecting overshoot peaks of the audio signal. Each frame of sound is then analyzed in the frequency domain. Basically, voices exhibit low-frequency content, thus they can be evaluated from low-frequency sound. The detected scene is then complemented by considering continuous voice energy, based on the assumption that a conversation does not have long silent periods. This process can detect conversation scenes by using only audio data. Therefore, the detected scenes contain all the sound from talking voices, as well as undesired noise. The preferred scenes can be extracted by the following process.

3.2 Human Face Extraction

This process is performed to detect face-to-face conversation scenes. Normally, people like to emphasize an important topic by making eye contact with the partner. Conversation scenes in video data should therefore contain the partner's face during the conversation period.

We transform video data to the hue saturation value (HSV) color space to detect human skin color based on the thresholds $H < 25/360$ and $H > 335/360$, $0.2 < S < 0.6$, and $V > 0.4$. Post-processing is then performed on the assumption that a face is present. Morphological operations including closing and opening tools are used to complete the shape of the face and remove any noisy pixels in the background. The size of the color face area is calculated on the key frames of detected scenes to decide whether the scenes contain a talking face. If the examined frame scores above the decision threshold, this scene will be classified as a conversation scene with an important topic.

3.3 Conversation Scene Detection

Conversation scenes can be retrieved based on audio/visual information. A conversation scene without consideration of eye contact can be detected based on only audio information. Moreover, important conversation scenes assumed as conversations with eye contact can be detected based on audio and visual information. Thus, scenes detected with a human face can be classified. Interesting conversation scenes can then be selected from the presented frames. Examples of detected scenes including conversations without eye contact and face-to-face conversations are demonstrated in Fig. 3.



Fig. 3. Examples of conversation scenes in a restaurant

The experimental results show that the interesting scenes both of face-to-face conversation and conversation scenes without face-to-face contact could be detected. All conversation scenes were retrieved based on only audio features. However, the face-to-face conversation scenes can be extracted by using both audio and visual features to increase the accuracy and acquire more specific scenes.

The conversation scene detection algorithm was applied to a personal recording video for the Life Log system including university, park, and lunchtime sequences. University and park sequences were captured when the user went to the destination with a friend. Video of lunchtime was captured with many friends in a restaurant.

Table 1 presents the evaluation of conversation scene detection, for all detected scenes and for face-to-face conversation scenes. The experiments demonstrated that the algorithm can work well for university, and park sequences as seen in their precision and recall above 80 percent. However, more false alarms and missed detection occurred in the lunchtime video. This video contains some noise and many talking people, so the algorithm could not detect voices effectively. The video also captured many faces during lunchtime. For these reasons, the possibility of efficient detection of the desired scenes was limited. However, we can identify more specific scenes and increase accuracy by using context information from the wearable sensors as mentioned in the following section.

Table 1. Evaluation of conversation scene detection

Video	Correct	False	Miss	Precision	Recall
University	6/4	1/0	1/1	0.86/1.00	0.86/0.80
Park	12/4	3/1	1/0	0.80/0.80	0.92/1.00
Lunchtime	12/5	4/2	3/2	0.75/0.71	0.80/0.71

A/B: (A) All conversation scenes, (B) Face conversation scenes.

4 Context Based Retrieval Algorithm

With our novel concept of video retrieval, we can use not only the content of the audiovisual data, but also context from wearable devices, including a brain wave analyzer, motion sensors, GPS receiver, databases and the computer operating system to support efficient retrieval system. The details of our contexts are provided as follows.

4.1 Brain Wave Analyzer

The brain wave signal named the α -wave was acquired from the brain wave analyzer and can show a person's arousal status. When the α -wave has a small signal or α -blocking is occurring, the person is considered to be aroused or to be interested in something. Our previous work demonstrated that an interesting scene based on the user's emotion can be effectively retrieved using brain waves [1]. The brain wave is thus a useful context for detecting when the user is interested. Conversation scenes can be detected this way, because a person must concentrate on the partner's conversation to understand it. Hence, this brain wave signal is useful in classifying conversation scenes where the user is paying attention.

4.2 Motion Sensors

We can acquire motion information for our activities from motion sensors. This information can identify activities such as walking, running, or not moving. The data from an acceleration sensor and a gyro sensor were processed as training data with K-Means and the HMM method to classify the user's movement [9]. When the user is moving, the acceleration sensor will be active and will give high-frequency signals. On the other hand, the acceleration sensor will give small signals when the user moves slowly or is stationary. Hence, we can use this information if we assume that we usually have important conversations when we move slowly or are stationary. This feature can be used to confirm the accuracy of scene detection.

4.3 GPS Receiver

We can acquire GPS data from the GPS receiver, and thereby know the user's position as longitude and latitude coordinates. The contents of videos and the

location information are recorded synchronously. The system can convert longitude and latitude into a postal address using the town database, and this can be used as a context key for video retrieval. Furthermore, we can use latitude and longitude information to identify the user's location by plotting the position as footprints on the town map. Hence, the concept of "my university" can identify the location where a conversation takes place, and can be used as a retrieval key.

4.4 Town Database

The system has a town database identifying stores, companies, restaurants, and other places in Japan. We can search for information about shops by their name or category and by their address relative to a location from the GPS data. Thus, a conversation scene in 'the sushi restaurant near my university', for example, can be retrieved based on the database of sushi restaurants close to the university.

4.5 Operating System Data

The portable computer's operating system also records the time synchronously. We can therefore acquire the present time associated with the audiovisual information of Life Log video. Hence, date and time can support this retrieval method directly from the operating system. We can thus retrieve a required conversation scene during a lunchtime based on the definition of the time period.

A display of useful context information from the Life Log system is shown in Fig. 4.

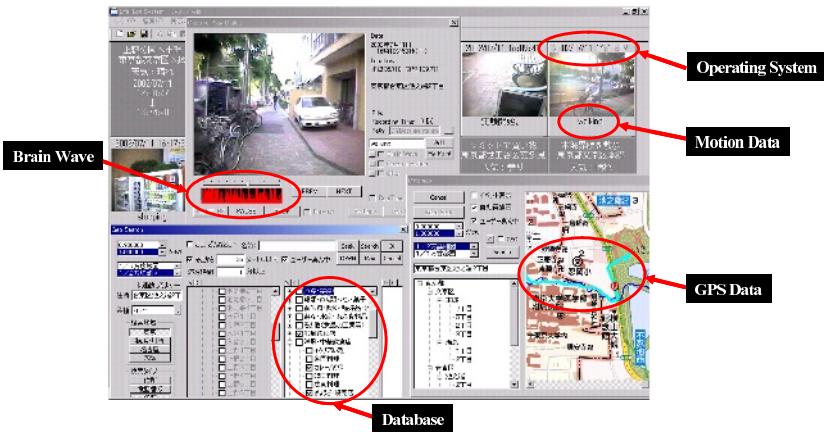


Fig. 4. Displaying context information from the Life Log system

5 The Concept of the Retrieval Method

As we mentioned above, in our novel retrieval method the content, context, and concept provide different pieces of important information. Hence, effective use of the information depends on the user's preference in retrieving desired scenes from our experiences.

5.1 The Combination of Contexts

To confirm the accuracy of retrieved scenes, many contexts are required to support a particular person's query, because if only one kind of key is used, too many scenes that satisfy a query will appear. Therefore, a combination of many different keys should isolate the desired result or at least eliminate most of the undesired scenes.

5.2 The Usage of Contents and Contexts

According to content and context-based features, it is possible to determine the significance of retrieval keys. For example, if we simply requested conversation scenes, in the first instance they will all be retrieved based on content features to observe all talking topics. On the other hand, if we are interested in a more specific location, the location from GPS as a context will be used as the first key to specify the exact place of the activity. Then we can identify the conversation scenes among the events at this exact place based on the contents of audiovisual data.

6 Summary

A novel approach to video retrieval in the Life Log application was explained to demonstrate the importance of content and context information. There are many significant features, from audiovisual data as content and data from various environmental devices as context that can be acquired from the Life Log system. Conversation scenes are an example of applying our novel concept to retrieval, based on low-level features of the audiovisual information. More specific conversation scenes can be retrieved based on additional context in the Life Log system. The integration of content and context features is very advantageous to support efficient Life Log retrieval. In the near future, we believe that research on this concept-based retrieval will continue and will exploit both low-level and high-level concepts to support the user's preferences and extend our viewpoint on multimedia retrieval.

References

1. Aizawa, K., Ishijima, K.: Summarizing Wearable Video. Proc. Intl. Conf. of ICIP 2001, Vol.3, pp. 398-401, Oct. 2001.

2. Smeulders, A.W.M., Worring, M., Santini, S.: Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.
3. Aslandogan, Y.A., Yu, C.T.: Techniques and Systems for Image and Video Retrieval. *IEEE Trans. on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 56-63, Jan.-Feb. 1999.
4. Pan, H., Liang, Z.P., Huang, T.S.: Fusing Audio and Visual Features of Speech. *IEEE Proc. Intl. Conf. of ICIP 2000*, pp. 214-217.
5. Cuetos, P., Neti, C.: Audio-Visual Intent-to-Speak Detection For Human-Computer Interaction. *IEEE Proc. Intl. Conf. of ICASSP 2000*, pp. 2373-2376.
6. Minami, K., Tonomura, Y.: Video Handling with Music and Speech Detection. *IEEE Multimedia*, pp. 17-25, July 1998.
7. Tancharoen, D., Jitapunkul, S.: Automatic Face Color Segmentation Based Rate Control For Low Bit Rate Video Coding. *Proc. Intl. Sym. of ISCAS 2003, Vol.2*, pp. 384-387, May 2003.
8. Tsekeridou, S., Pitas, I.: Content-Based Video Parsing and Indexing Based on Audio-Visual Interaction. *IEEE Trans. Circuits and System for Video Technology*, Vol. 11, No. 4, April 2001.
9. Sawahata, Y., Aizawa, K.: Wearable Imaging System for Summarizing Personal Experiences. *Proc. Intl. Conf. of ICME 2003*.
10. Hori, T., Aizawa, K.: Context-Based Video Retrieval System for the Life-log Applications. *Proc. Intl. Conf. of MIR 2003, ACM*, pp.31-38.

Content-Based Retrieval of 3D Head Models Using a Single Face View Query

Pui Fong Yeung, Hau San Wong, and Horace H.-S. Ip

Image Computing Group, Department of Computer Science,
City University of Hong Kong
pfyung@cs.cityu.edu.hk, {cshswong, cship}@cityu.edu.hk

Abstract. In this paper, we propose a new approach for 3D human head model retrieval using a single 2D face view only. In this way, the query can be conveniently specified in the form of a single portrait which is in most cases readily available, instead of a 3D model query which is difficult to construct, or text-based query which in general cannot describe the models adequately. To realize this approach, a mapping between the 2D face views and the 3D models needs to be established. In our case, 3D models are represented with a set of adaptive basis functions, while their corresponding 2D face views are characterized with a set of eigenface basis functions. In this way, a particular model and its associated face view can be identified by two separate set of expansion coefficients. To associate the two, we propose to exploit neural network techniques to identify a mapping. With this 2D-3D mapping, we can thus estimate a set of associated 3D expansion coefficients for the input query to retrieve the relevant models in the database.

Keywords: Computer graphics, 3D model retrieval, neural networks.

1 Introduction

With the wide availability of 3D scanners and digitizers, construction of 3D computer graphics models is no longer an arduous task, and one can readily build up a large collection of 3D models in game design, manufacturing, and computer animation. The availability of these large collections in turn implies the increasing need to develop effective content-based retrieval techniques for these models.

In previous works, 3D model retrieval is performed based on text, keywords or even by a specially designed language [1]. While these approaches are easy to implement, they are not applicable to the retrieval of 3D head model as it is difficult to describe facial features of a human head model and to identify a unique head model by words.

Other approaches use color, texture and shape information [2,3,4,5,6,7,8,9, 10] for characterizing the models, with some of these features designed such that they are invariant to rotation, scaling and reflection [11]. While these features are useful for identifying classes of models with distinct characteristics, they

may not be adequate for characterizing the subtle differences between the facial features of different head model classes.

This problem is alleviated to a certain extent by recently proposed indexing approaches. In [12], a hierarchical indexing scheme was proposed for 3D head models based on facial region similarity. In [13], a clustering technique based on hierarchical SOM is adopted for model indexing, while in [14], evolutionary computation is applied to optimize the feature representation of the models. However, we still need to present the query in the form of a 3D model.

In view of these problems, we propose a novel approach for 3D head model retrieval which uses *a single 2D face view* as query. In this way, the query can be conveniently specified in the form of a single portrait which is in most cases readily available, instead of a 3D model query which is difficult to construct, or text-based query which in general cannot describe the models adequately. To realize this approach, we specify the human head shapes in the form of a set of adaptive basis functions constructed from a set of 3D head models. In this way, the space of possible human head shapes can be characterized, and a particular model in this space can also be identified by its set of expansion coefficients. In a similar way, their corresponding 2D face views are characterized with a set of eigenface basis functions [17].

Given these representations, we then need to establish an accurate 2D-3D mapping between the two sets of expansion coefficients. While various forms of linear mappings were proposed to represent this correspondence in the related task of 3D model reconstruction in [15,16], they may not be able to capture the complexity of the association between the 2D and 3D spaces.

To overcome this problem, we propose to adopt neural network techniques to approximate the 2D-3D association. The reason for adopting neural networks is that they are capable of representing the complex and non-linear mapping between these two spaces due to their universal approximation capability. In addition, the mapping itself can be conveniently constructed based on a set of training examples through a suitable learning algorithm, instead of requiring the pre-determination of a specific functional form for the mapping.

2 Model and Face View Representation

In our approach, the set of 3D models are represented by a function $\mathbf{f}(\mathbf{z})$, where \mathbf{z} is the position vector associated with a suitable parameterization of the model surface. Each position vector is then associated with a surface property, such as the corresponding 3-D coordinates of the surface point, the local curvature, or the surface normal direction, through the mapping \mathbf{f} . Our task is to find a set of basis functions $\mathbf{g}_1(\mathbf{z}), \dots, \mathbf{g}_N(\mathbf{z})$, such that each model can be adequately represented by this function set. In our approach, we achieve this by applying principal component analysis (PCA) and adopting the resulting set of eigenfunctions for characterization. Specifically, the coordinates of the vertices of each model are concatenated to form a single vector. Through PCA, a set of eigenvectors is extracted and each model can be specified by the set of projection

coefficients $\mathbf{v} = [v_1, \dots, v_N]^T$ on the eigenvectors. In other words, each model can be represented as follows

$$\mathbf{f}(\mathbf{z}) = \sum_{n=1}^N v_n \mathbf{g}_n(\mathbf{z}) \quad (1)$$

We next extract a set of features from the 2-D face views to form an association with the 3-D basis coefficients. Specifically, for a set of 2-D face views corresponding to the 3-D head model training set, we can derive a set of eigenface functions $s_1(\mathbf{x}), \dots, s_M(\mathbf{x})$, such that each face view $r(\mathbf{x})$ can be represented by these eigenfaces as follows

$$r(\mathbf{x}) = \sum_{m=1}^M u_m s_m(\mathbf{x}) \quad (2)$$

where \mathbf{x} is the position vector on the 2D image lattice. A face view can then be described by a set of eigenface projection coefficients $\mathbf{u} = [u_1, \dots, u_M]^T$ while its corresponding model can be represented by $\mathbf{v} = [v_1, \dots, v_N]^T$.

We then determine an optimal mapping $\mathbf{h} : \mathbf{R}^M \rightarrow \mathbf{R}^N$ such that the following condition is valid:

$$\mathbf{v} \approx \mathbf{h}(\mathbf{u}) \quad (3)$$

In other words, for a set of P 3-D models with representations \mathbf{v}_p , $p = 1, \dots, P$, we search for an optimal mapping \mathbf{h} such that the following error measure E_h is minimized

$$E_h = \sum_{p=1}^P \|\mathbf{v}_p - \mathbf{h}(\mathbf{u}_p)\|^2. \quad (4)$$

We implement the mapping \mathbf{h} between the two sets of expansion coefficients in the form of a neural network. As discussed previously, we choose neural network for the mapping due to its universal approximation capability and the possibility of adaptive construction of this mapping through a set of training examples

For a given 2D face view query \mathbf{u}_q , we can obtain the associated set of 3D basis coefficients through the previously constructed mapping as follows:

$$\mathbf{v}_q = \mathbf{h}(\mathbf{u}_q) \quad (5)$$

After obtaining the set of 3D projection coefficients, we can then compare these with the set of coefficients \mathbf{v} associated with each model in our database based on the Euclidean distance $\|\mathbf{v}_q - \mathbf{v}\|$. These distance measures are then ranked in ascending order to produce the retrieved model list.

3 Experimental Result

Our proposed method is applied to a database containing 1000 3D head models categorized into 10 classes, which is constructed based on the MPI Face Database [18]. Each model consists of 6152 polygons with 6293 vertices.

To construct the training set, we extract the 2D front view projection of each model, followed by applying eigen-analysis on these grayscale images to obtain a set of eigenfaces. For the 3D models, we extract a set of eigenfunctions by performing eigen-decomposition on the entire dataset. We observe that the first 10 eigenfunctions are adequate for model feature characterization for 3D head models, while 2D face views require the first 20 eigenfaces in order to give an accurate description.

The eigenface projection coefficients for a face view and its corresponding 3D eigenfunction coefficients are then considered as a training pair for the feedforward neural network which has a single hidden layer with 60 nodes. Gradient descent learning is then applied to the network until the mean squared error is below a prescribed threshold.

To evaluate the performance of the proposed retrieval approach, 10 sample queries not in the training set are presented to the network to obtain their associated 3D projection coefficients. These are then compared with the corresponding coefficients associated with the database models based on the Euclidean distance. By ranking these distance measures, we can obtain a 3D model retrieval list.

The retrieval performance based on 2D queries is summarized by the (solid) precision-recall curve in Fig. 1. It can be observed that the precision level remains high for a large range of recall values, as indicated by the slowly decreasing slope

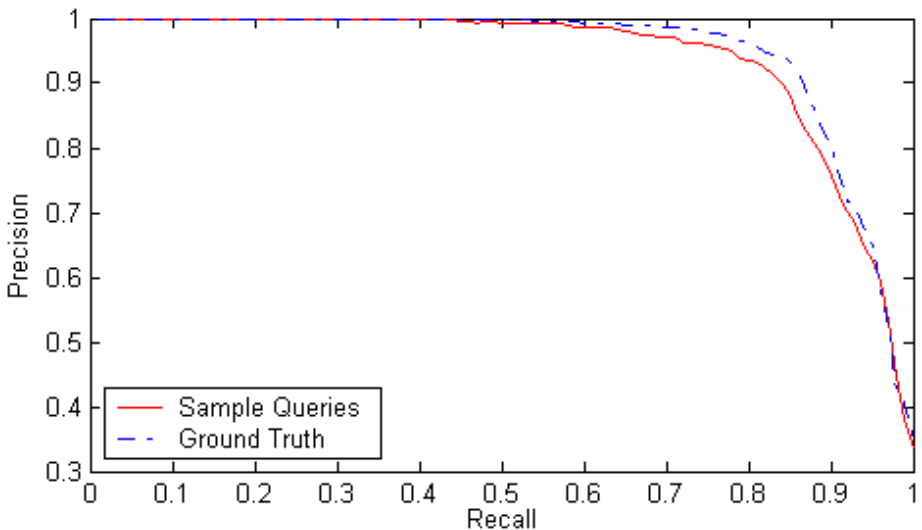






Fig. 1. Precision recall characterization of retrieval performance

Table 1. Sample queries and the retrieved models

Sample Query	Retrieved Models					
						
						

of the curve, which implies that most of the relevant models are at the top of the retrieval list. The performance can be compared with the case where the actual 3D projection coefficients of the queries, instead of their approximations based on the neural mapping, are directly used for retrieval (dashed curve). The proximity of the two curves indicates that the neural network is capable of accurately representing the mapping between the 2D and 3D projection coefficients.

To further illustrate the retrieval results, we have displayed the first 5 retrieved models for 2 sample queries in Table 1. It can be observed that the retrieved models bear a close resemblance to the queries in each case.

4 Conclusion

In this paper, a 3D human head model retrieval approach based on a single 2D face view image is proposed. In our approach, we represent a particular 3D model by a set of eigenfunctions while each face view image is characterized by a set of eigenfaces. The mapping between these two sets of coefficients is represented in the form of a single-hidden-layer neural network. Our experiment shows that the proposed approach is capable of retrieving 3D models which bear a close resemblance to the 2D query, as indicated by the resulting precision-recall curves

and visual observations. More importantly, instead of a complete 3D model, the query can now be conveniently specified in the form of a single portrait which can be readily captured due to the wide availability of web cameras.

Acknowledgement. The work described in this paper was partially supported by a grant from the Research Grants Council of Hong Kong Special Administrative Region, China [Project No. CityU 1197/03E] and a grant from City University of Hong Kong [Project No. 7001596]. Part of the dataset used in the experiments was provided by the Max Planck Institute for Biological Cybernetics in Tuebingen, Germany.

References

1. Horikoshi, T., Kasahara, H., "3-D shape indexing language", Proc. 9th Annual Intl. Phoneix Conf. on Computers and Communications, pp. 493-499, 1990
2. Jobst Loffler, "Content-based Retrieval of 3D Models in Distributed Web Databases by Visual Shape Information", Proc. International Conference on Information Visualisation (IV2000), p.82-87, London, England, 19 - 21 Jul, 2000
3. Eric Paquet and Marc Rioux, "Content-Based Access of VRML Libraries", Proc. IAPR Intl. Workshop on Multimedia Information Analysis and Retrieval, Hong Kong, China, pp. 20-32, Aug 1998
4. E. Paquet, M. Rioux, A. Murching, T. Naveen and A.Tabatabai, "Description of Shape Information for 2-D and 3-D Objects", Signal Processing: Image Communication, vol.16, 103-122 (2000)
5. R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Matching 3D Models with Shape Distributions", Proc. Intl. Conf. on Shape Modelling and Applications, SMI 2001, Genova, Italy, pp. 154-166, May 2001
6. D. Zhang and T. Chen, "Efficient Feature Extraction for 2D/3D Objects in Mesh Representation", Proc. Intl. Conf. Image Processing, Thessaloniki, Greece, pp. 935-938, Oct 2001.
7. D. V. Vranic, D. Saupe, and J. Richter, "Tools for 3D-object Retrieval: Karhunen-Loeve Transform and Spherical Harmonics", Proc. IEEE 2001 Workshop Multimedia Signal Processing, Cannes, France, pp. 293-298, Oct 2001
8. D. Saupe and D. V. Vranic, "3D Model Retrieval with Spherical Harmonics and Moments", Proc. DAGM 2001, Munich, Germany, pp. 392-397, Sep 2001
9. R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape Distributions", ACM Trans. on Graphics, 21(4), 807-832 (2002)
10. T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman and D. Dobkin, "A Search Engine for 3D Models", ACM Trans. on Graphics, 21(3), 83-105 (2003).
11. Dejan V. Vranic, "Content-Based Search for 3D-Objects", Proc. Fourth International Conference on Computational Intelligence and Multimedia Applications (IC-CIMA'01), pp. 266-270, Yokusike City, Japan, 30 Oct - 1 Nov 2001.
12. H. H. S. Ip and W. Y. F. Wong, "3D Head Models Retrieval Based on Hierarchical Facial Region Similarity", Proc. Int. Conf. on Vision Interface (VI 2002), pp. 314-319, 2002.
13. Hau-San Wong, Kent K. T. Cheung , Yang Sa, Horace H. S. Ip, "Indexing and Retrieval of 3D Models by Unsupervised Clustering with Hierarchical SOM", accepted for ICPR 2004, Cambridge, United Kingdom, 23-26 Aug 2004

14. Kent K. T. Cheung, Hau-San Wong, Horace H. S. Ip, "3D Graphical Model Retrieval by Evolutionary Computation-based Adaptive Histogram Binning", Proc. DMS 2003, pp. 420-425, Miami, USA, 24-26 Sep 2003
15. V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces", Proc. Siggraph 99. pp. 187-194, 1999
16. V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model", IEEE Trans. on PAMI, vol. 25, no. 9, pp. 1063-1074, 2003
17. M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cognitive Neuroscience, vol. 3, pp. 71-86, 1994
18. MPI Face Database: <http://faces.kyb.tuebingen.mpg.de/>

Region-Based Image Retrieval with Perceptual Colors

Ying Liu¹, Dengsheng Zhang¹, Guojun Lu¹, and Wei-Ying Ma²

¹ Gippsland School of Computing and Information Technology,
Monash University, Vic, 3842, Australia,

{ying.liu, dengsheng.zhang, guojun.lu}@infotech.monash.edu.au,

² Microsoft Research Asia, No. 49 ZhiChun Road, Beijing, 100080, China,
wyma@microsoft.com

Abstract. Due to the ‘semantic gap’ between low-level visual features and the rich semantics in user’s mind, performance of traditional content-based image retrieval systems is far from user’s expectation. In attempt to reduce the ‘semantic gap’, this paper introduces a region-based image retrieval system with high-level semantic color names used. For each segmented region, we define a perceptual color as the low-level color feature of the region. This perceptual color is then converted to a semantic color name. In this way, the system reduces the ‘semantic gap’ between numerical image features and the richness of human semantics. Four different ways to calculate perceptual color are studied. Experimental results confirm the substantial performance of the proposed system compared to traditional CBIR systems.

1 Introduction

To overcome the drawback of traditional text-based image retrieval systems which require considerable amount of human labors, content-based image retrieval (CBIR) was introduced in the early 1990’s. CBIR indexes images by their low-level features, such as color, shape, texture. Commercial products and experimental prototype systems developed in the past decade include QBIC system[1], Photobook system[2], Netra system[3], SIMPLIcity system [4], etc. However, extensive experiments on CBIR systems show that in many cases low-level image features can’t describe the high level semantic concepts in the user’s mind. Hence, the performance of CBIR is still far from the user’s expectations [5][6]. ‘The discrepancy between the relatively limited descriptive power of low-level imagery features and the richness of user semantics’, is referred to as the ‘semantic gap’ [7].

In order to improve the retrieval accuracy of CBIR systems, research focus in CBIR has been shifted from designing sophisticated feature extraction algorithms to reducing the ‘semantic gap’[8]. Recent work in narrowing down the ‘semantic gap’ can be roughly classified into 3 categories: 1) Using region-based image retrieval (RBIR) which represents images at region-level with the intention to be more close to the perception of human visual system [9]. 2) Introducing

relevance feedback into image retrieval system for continuous learning through on-line interaction with users to improve retrieval accuracy [7]. 3) Extracting semantic features from low-level image features using machine learning or data mining techniques [5].

We intent to develop a RBIR system with high-level concepts obtained from numerical region features such as color, texture, spatial position. This paper includes our initial experimental results using semantic color names. Firstly, each database image is segmented into homogeneous regions. Then, for each region, a perceptual color is defined. This is different from conventional methods using color histogram or color moments [4][9]. The perceptual color is then converted to a semantic color name (for example, 'grass green', 'sky blue'). In this way, the 'semantic gap' is reduced. Another advantage of the system is that it allows users to perform query by keywords (for example, 'find images with sky blue regions').

The remaining of the paper is organized as follows. In section 2, we describe our system in details. Section 3 explains the test data set and the performance evaluation model. Experimental results are given in Section 4. Finally, Section 5 concludes this paper.

2 System Description

Our system includes three components, image segmentation, color naming and query processing.

2.1 Image Segmentation

Natural scenes are rich in both color and texture, and a wide range of natural images can be considered as a mosaic of regions with different colors and textures. We intent to relate low-level region features to high-level semantics such as color names used in daily life (pink, green, sky blue, etc), real-world texture patterns (grass, sky, trees, etc). For this purpose, firstly we use 'JSEG'[10] to segment images into regions homogeneous in color and texture. Fig.1 gives a few examples.

2.2 Color Naming

Perceptual Colors: In stead of using traditional color features such as color moments or color histograms[4][9], we define a perceptual color for each segmented region with the intention to relate it to semantic color names.

Although millions of colors can be defined in computer system, the colors that can be named by users are limited [11]. For example, the first two colors in Fig. 2 correspond to two different points in HSV (Hue, Saturation, Value) space, but users are likely to name them both as 'pink'. Similarly, both of the next 2 colors could be named as 'sky blue'. The HSV values (with ranges [0,360], [0,100], [0,100], respectively) of the 4 colors are given below.

Pink: (H,S,V) = (326, 42,100), (330, 40, 100)

SkyBlue: (H,S,V) = (200, 42, 93), (202, 40,100)

HSV color space is the most natural color space in visual. We define a perceptual color in HSV space for each region and then convert it to a semantic color name. Four different ways to define perceptual color are studied.

- We use the average HSV value of all the pixels in a region as its perceptual color (referred to as ‘Ave-cl’). This is reasonable as most regions obtained using JSEG are color homogeneous.
- The value of Hue is in circular, for example, both ‘0’ and ‘360’ represents ‘red’ color. Averaging Hue values may result in a color very different from what we expect. For example, $(0+360)/2=180$, this means the average of two ‘red’ pixels is ‘cyan’. To solve this problem, we first calculate the average RGB value of a region and then convert it to HSV domain. This result is referred to as ‘RGB-cl’.
- Due to the inaccuracy in image segmentation, pixels not belonging to the interested region might be included in ‘Ave-cl’ calculation and results in a color perceptually different from that of the region. Hence, we consider using the dominant color of a region as the perceptual color. For this, we first calculate the color histogram ($10*4*4$ bins) of a region and select the bin with maximum size. The average HSV value of all the pixels in the selected bin is used as the dominant color and referred to as ‘Dm-cl’.
- Considering that the histogram of a region may contain more than one bins of large size, we calculate the average HSV value of all the pixels from $M(i,1)$ large bins as the perceptual color. Experimentally, we select all those bins with size no less than 68% of the maximum-size bin. The result is referred to as ‘Dmm-cl’.

We observed that in most cases the four perceptual colors are very similar, as in Fig. 3(1), 3(2). However, in some special cases, ‘Ave-cl’ results in a color visually very different from that of the original region. For example, in region 3(a), due to the inaccuracy in segmentation, a small part of the green background (left side) is included in the flower. In addition, some pixels are not of pink color, but dark yellow (at the center of the flower) or gray (in between the petals). The result is that the ‘Ave-cl’ in 3(b) turns to be different from the color of region in 3(a).

Color Naming: Color naming is to map a numerical color space to semantic color names used in natural language. Qualification color naming model is often used, in which Hue value is quantized into a small set of about 10-20 base color names [12]. In [12], the author uniformly quantized the Hue value into 10 base colors, such as red, orange, yellow, etc. Saturation and Luminance are quantized into 4 bins respectively as adjectives signifying the richness and brightness of the color. There are two problems with the model used in [14]. Firstly, uniform quantization of Hue value is not proper as colors in the HSV space are not uniformly distributed (refer to Fig. 4). The reason is that different colors have different wave bandwidths. For example, the wave band of yellow and blue are 565-590nm, 450-500nm, respectively. The second problem is that in [12], ‘red’

corresponds to Hue value from 0 to 36 (normalized to 0-0.1 in [12]). However, we notice that Hue of ‘red’ can be around either 0 or 360.

Considering the above mentioned problems, we design a color naming model as follows. Firstly, we define 8 base colors, red, orange, yellow, green, cyan, blue, purple, magenta, with the range of the Hue values as [0,8) or [345,360], [8,36), [36,80), [80,160), [160,188), [188,262), [262,315), [315,345), respectively. Saturation and Value are quantized into 3 bins as in Fig. 5, with the corresponding adjectives shown in Table 1. The asterisks indicate special cases. When $S=0$ and $V=1$, we have ‘grey’. When $S=0$ and $V>80$, we have ‘white’. When $V=0$, we always get ‘black’. Base color names with their adjectives can be simplified as other common-used color names. For instance, ‘pale magenta’ is named as ‘pink’.

Finally, we obtain $8*2*2+3=35$ different colors. For example, the first two colors in Fig. 2 are both named as ‘pink’. Similarly, the other two colors are named as ‘sky blue’.

In this way, the low-level color features are mapped to high-level semantic color names, thus reducing the ‘semantic gap’.

2.3 Query Processing

All database images are segmented into regions and their low-level color features and color names are stored for retrieval purpose. The system can support different types of queries.

1) Query by specified region - The user selects an interested region from an image as the query region. The system calculates the low-level color feature and color name of the query region. All images containing region(s) of same color name are selected and form a candidate set C . Then, the images in C are further ranked according to their EMD[13] distance to the query image. With region distance defined as the Euclidean distance between region color features, EMD measures the overall distance between two images.

2) Query by keyword - The keyword is selected from the 35 semantic colors defined. In this case, the system returns all images containing region(s) of same color name as specified by the keyword.

In this paper, we work on the first case, which is more complex.

3 Database and Performance Evaluation

Corel data set is often used to evaluate the performance of image retrieval systems due to its large size, heterogeneous content and human annotated ground truth available. However, to be used in image retrieval system as test set, some pre-processing work is necessary for the following two reasons: 1) some images with similar content are divided into different categories. For examples, the images in ‘Ballon1’ and ‘Ballon2’. 2) Some ‘category labels’ are very abstract and the images within the category can be largely varied in content. For instance, the category ‘Australia’ includes pictures of city building, Australian wild animals, etc. A few examples are given in Fig. 6.

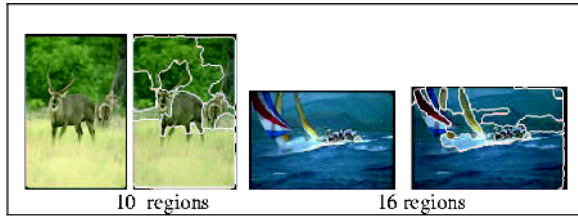


Fig. 1. JSEG segmentation results



Fig. 2. Example colors

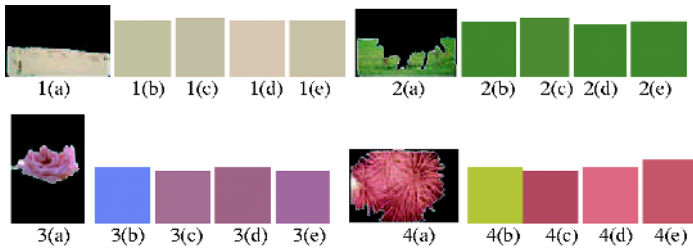


Fig. 3. Region perceptual colors (a) original region, (b) ‘Ave-cl’, (c) ‘RGB-cl’, (d) ‘Dm-cl’, (e) ‘Dmm-cl’



Fig. 4. HSV color space (H,S)

If Saturation>55 S=2;
 else if Saturation>8 S=1;
 else S=0;
 If Value>58 V=2;
 else if Value>15 V=1;
 else V=0;

Fig. 5. Quantization of S,V

Table 1. Adjectives according to S,V

	0	1	2
S	*	Pale	Normal
V	*	Dark	Normal



Fig. 6. Example images from category ‘Australia’

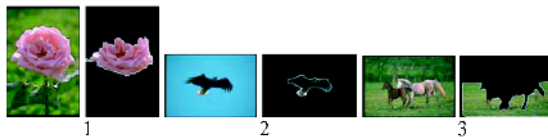


Fig. 7. Query images/regions examples

Hence, it's better to select a subset from Corel images with ground truth data available, or make some necessary changes in setting the group truth data.

We selected 5,000 Corel images as our test set (ground truth available). 'JSEG' segmentation produces 29187 regions (5.84 regions per image on average) with size no less than 3% of the original image. We ignore small regions considering that regions should be large enough for us to study their texture patterns later.

Precision and recall are often used in CBIR system to measure retrieval performance. Precision (Pr) is defined as the ratio of the number of relevant images retrieved N_{rel} to the total number of retrieved images N . Recall (Re) is defined as the number of relevant images retrieved N_{rel} over the total number of relevant images available in the database N_{all} . We calculate the average Pr and Re of 30 queries with $N=10,20,\dots,100$, and obtain the Pr~Re curve. A few query images and the specified regions are displayed in Fig. 7.

4 Experimental Results

Firstly, we compare the performance of our RBIR system using 'Ave-cl', 'RGB-cl', 'Dm-cl' and 'Dmm-cl' respectively. The Pr~Re curves are given in Fig.8(a). The results show that 'Dm-cl' and 'Dmm-cl' perform better than 'Ave-cl' does. 'RGB-cl' works better than 'Ave-cl' but not as good as 'Dm-cl' and 'Dmm-cl'. In addition, the performance of 'Dm-cl' is very close to that of 'Dmm-cl'. In this work, we use 'Dmm-cl'. Fig.9 compares the retrieval results for query 1 using 'Ave-cl' and 'Dmm-cl'.

Our experiments also show that the proposed color naming system works better than that used in [12]. Due to space limitation, we did not give the results here.

In addition, we compare our system (denoted as 'R') with a CBIR system using global color histogram (referred to as 'G'). In system 'G', images are represented by their HSV space color histogram with H, S, V uniformly quantized into 18, 4, 4 bins, respectively. The similarity of two images is measured by the Euclidean distance between their color histograms.

We observed that 'R' works well when the interested region is recognized and the color names defined can well describe it. For example, in query 2, the query region is the 'eagle'. 'R' recognizes 'eagle' and successfully finds many relevant images. Fig.10 gives the retrieval results, with 'R' returns 8 relevant images within the top 10 retrieved, while 'G' finds only 3.

In another case, such as query 3, both 'R' and 'G' work well. Due to the large green background available in the query image and the relevant database images, retrieval accuracy of 'G' is very high. On the other hand, color name 'grass green' can well represent the grass region. Hence, retrieval performance of 'R' is also very good. Among the first 10 images retrieved, the number of relevant images retrieved by 'G' and 'R' are both 10.

Fig.8(b) compares the performance of 'G' and 'R' over 30 queries.

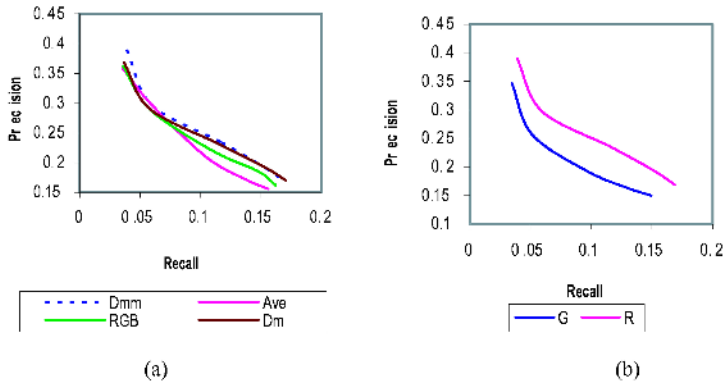


Fig. 8. (a) Using different perceptual colors, (b) ‘G’-‘R’ over 30 queries



Fig. 9. Retrieval Results for query 1. The first image is the query image. ‘Q’ refers to query region. ‘T’ refers to the relevant images selected.

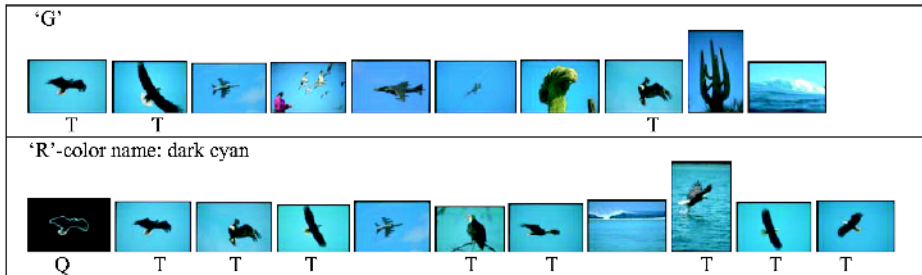


Fig. 10. Retrieval Results for query 2. The first image is the query image. ‘Q’ refers to query region. ‘T’ refers to the relevant images selected.

5 Conclusions

This paper presents a region-based image retrieval system using high-level semantic color names. For each segmented region, a perceptual color is defined,

which is then converted to a semantic color name using our color naming algorithm. In this way, the system reduces the ‘semantic gap’ between numerical image features and the richness of human semantics. Experimental results confirm the substantial performance of the proposed system over conventional CBIR systems.

In our future work, we will make use of multiple types of low-level image features to extract more accurate semantics. We expect the performance of our system to be further improved.

References

1. C.Faloutsos, R.Barber, M.Flickner, J.Hafner, W. Niblack, D.Petkovic, and W.Equitz, “Efficient and Effective Querying by Image Content,” *J. Intell. Inform. Syst.*, vol.3, no.3-4, pp231-262,1994
2. A. Pentland, R.W.Picard, and S.Scaroff, “Photobook: Content-based Manipulation for Image Databases”, *Inter. Jour. Computer Vision*, vol. 18, no.3, pp233-254, 1996.
3. W.Y.Ma and B.Manjunath, “Netra: A Toolbox for Navigating Large Image Databases”, *Proc. of ICIIP*, pp568-571, 1997.
4. J.Z.Wang, J.Li, and G. Wiederhold, “SIMPLiCity: Semantics-Sensitive Integrated Matching for Picture Libraries,” *IEEE Trans. Pattern and Machine. Intelligence*. Vol 23, no.9, pp947-963, 2001.
5. A.Mojzilovic, B.Rogowitz, “Capturing Image Semantics with Low-Level Descriptors”, *Proc. of ICIIP*, pp18-21, 2001
6. X.S. Zhou, T.S.Huang, “CBIR: From Low-Level Features to High-Level Semantics”, *Proc. SPIE Image and Video Communication and Processing*, San Jose, CA. Jan.24-28, 2000.
7. Yixin Chen, J.Z.Wang, R.Krovetz, “An Unsupervised Learning Approach to Content-based Image Retrieval”, *IEEE Proc. Inter. Symposium on Signal Processing and Its Applications*, pp197-200, July 2003.
8. Arnold W.M. Smeulders, Marcel Worring, Amarnath Gupta, Ramesh Jain, “Content-based Image Retrieval at the End of the Early Years”, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, No.12, Dec. 2000.
9. Feng Jing, Mingjing Li,, Lei Zhang, Hong-Jiang Zhang, Bo Zhang, “Learning in Region-based Image Retrieval”, *Proc. Inter. Conf. on Image and Video Retrieval(CIVR2003)*, 2003.
10. Y.Deng, B.S.Manjunath and H.Shin “Color Image Segmentation”, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR ’99*, Fort Collins, CO, vol.2, pp.446-51, June 1999.
11. E.B.Goldstein, *Sensation and Perception*, 5th Edition, Brooks/Cole, 1999.
12. Conway, D.M., “An Experimental Comparison of Three Natural Language Color Naming Models”, *Proc. East-West International Conference on Human-Computer Interactions*, St. Petersburg, Russia, pp328-339, 1992.
13. Rubner, Y., Tomasi, C., and Guibas, L., “A Metric for Distributions with Applications to Image Databases”, *Proc. of the 1998 IEEE Inter. Conf. on Computer Vision*, Jan. 1998.

Converting DCT Coefficients to H.264/AVC Transform Coefficients

Jun Xin, Anthony Vetro, and Huifang Sun

Mitsubishi Electric Research Laboratories, Cambridge MA 02139, USA
{jxin,avetro,hsun}@merl.com

Abstract. Many video coding schemes, including MPEG-2, use a Discrete Cosine Transform (DCT). The recently completed video coding standard, H.264/AVC, uses an integer transform, which will be referred to as HT in this paper. We propose an efficient method to convert DCT coefficients to HT coefficients entirely in the transform domain. We show that the conversion is essentially a 2D transform. We derive the transform kernel matrix, provide a fast algorithm and an integer approximation of the transform. We show that the proposed transform domain conversion outperforms the conventional pixel domain approach. It is expected to have applications in transform domain video transcoding.

Keywords: Transform domain video transcoding, video transcoding, H.264/AVC, MPEG-2.

1 Introduction

The transform used in many video coding schemes, including MPEG-2 [1], MPEG-1, and H.263 etc., is a Discrete Cosine Transform (DCT). The recently completed video coding standard, H.264/AVC [2], uses a low-complexity integer transform, hereinafter referred to as HT.

One important step in the transcoding of video from MPEG-2 format to H.264/AVC format is to convert the coefficients from the DCT domain to the HT domain, i.e. DCT-to-HT conversion. Fig. 1 shows the DCT-to-HT conversion in the context of intra-frame video transcoding.

Fig. 2 shows a pixel domain implementation of the DCT-to-HT conversion. The input is an 8×8 block (X) of DCT coefficients. An inverse DCT (IDCT) is applied to X to recover an 8×8 pixel block (x). The 8×8 pixel block is divided evenly into four 4×4 blocks (x_1, x_2, x_3, x_4). Each of the four blocks is passed to a corresponding HT to generate four 4×4 blocks of transform coefficients (Y_1, Y_2, Y_3, Y_4). The four blocks of transform coefficients are combined to form a single 8×8 block (Y). This is repeated for all blocks of the video.

It is desired to perform the transcoding entirely in the compressed or transform domain, then reconstructing the image pixels is avoided. Transform domain transcoding could be more efficient than the pixel domain transcoding because complete decoding and reencoding are not required [3]. Therefore, there is a

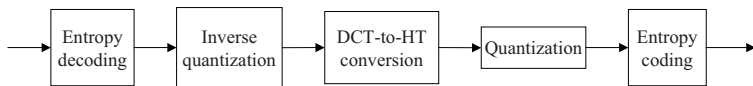


Fig. 1. Intra transcoding

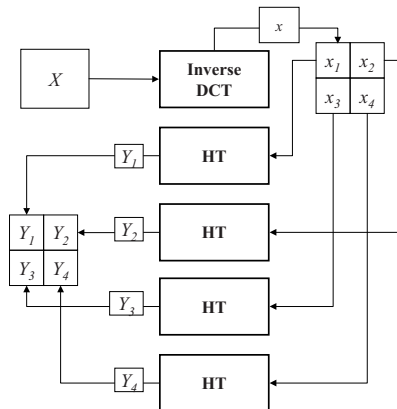


Fig. 2. Pixel domain HT-to-DCT conversion

need to have an efficient method to perform the DCT-to-HT conversion in the transform domain.

The paper is organized as follows. The proposed transform domain DCT-to-HT conversion is presented in Section 2. Section 3 and Section 4 discuss the fast algorithm and the integer approximation for the conversion respectively. Simulation results are given in Section 5. Section 6 concludes this paper.

2 DCT-to-HT Conversion

Fig. 3 shows our proposed transform domain DCT-to-HT conversion. In this paper, this conversion shall be called *S-transform*. It may be applied to the input DCT coefficients (X) of an input video in the MPEG-2 format to produce output HT coefficients (Y) of an output video in the AVC format. The *S-transform* is characterized by a transform kernel matrix, S , which is an 8×8 matrix:

$$Y = SX S^T \tag{1}$$

where S^T is the transpose of S . We shall derive S in the following.

The HT of x_1, x_2, x_3 and x_4 are $Y_1, Y_2, Y_3,$ and Y_4 respectively (Fig. 2), i.e.

$$\begin{aligned} Y_1 &= Hx_1H^T \\ Y_2 &= Hx_2H^T \\ Y_3 &= Hx_3H^T \\ Y_4 &= Hx_4H^T \end{aligned} \tag{2}$$

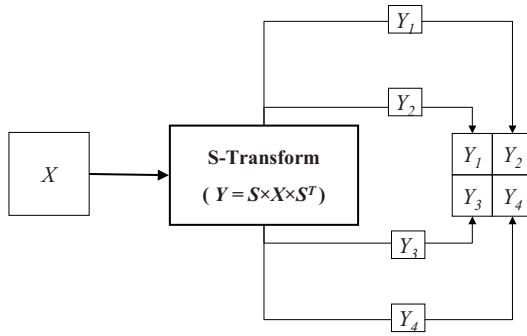


Fig. 3. Transform domain DCT-to-HT conversion

where H is the transform kernel matrix of HT: $H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{pmatrix}$

If $K = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}$, then we can rewrite (2) into a single equation

$$Y = K \times x \times K^T \tag{3}$$

where x is the IDCT of X . Let T_8 be the transform kernel matrix of DCT, we have $x = T_8^T X T_8$. It then follows that

$$Y = K \times T_8^T \times X \times T_8 \times K^T \tag{4}$$

Comparing (4) with (1), we have

$$S = K \times T_8^T \tag{5}$$

Therefore, the direct DCT-to-HT transform is given by (1) and its transform kernel matrix S , is

$$S = \begin{pmatrix} a & b & 0 & -c & 0 & d & 0 & -e \\ 0 & f & g & h & 0 & -i & -j & k \\ 0 & -l & 0 & m & a & n & 0 & -o \\ 0 & p & j & -q & 0 & r & g & s \\ a & -b & 0 & c & 0 & -d & 0 & e \\ 0 & f & -g & h & 0 & -i & j & k \\ 0 & l & 0 & -m & a & -n & 0 & o \\ 0 & p & -j & -q & 0 & r & -g & s \end{pmatrix} \tag{6}$$

where the values $a \dots s$ are (rounded off to four decimal places)

$$\begin{aligned} a &= 1.4142, & b &= 1.2815, & c &= 0.45, & d &= 0.3007, & e &= 0.2549, \\ f &= 0.9236, & g &= 2.2304, & h &= 1.7799, & i &= 0.8638, & j &= 0.1585, \\ k &= 0.4824, & l &= 0.1056, & m &= 0.7259, & n &= 1.0864, & o &= 0.5308, \\ p &= 0.1169, & q &= 0.0922, & r &= 1.0379, & s &= 1.975. \end{aligned}$$

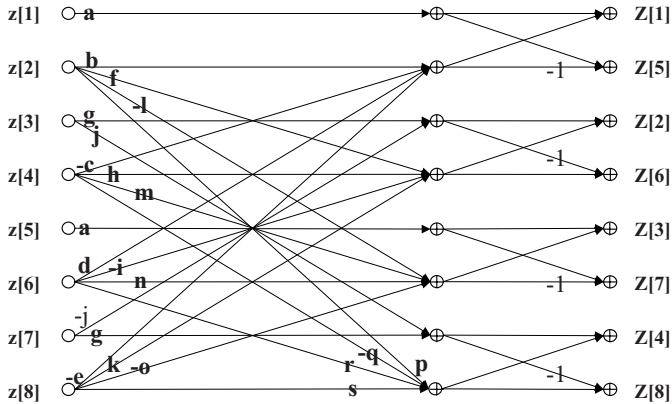


Fig. 4. Fast algorithm for the transform domain DCT-to-HT conversion

3 Fast DCT-to-HT Conversion

The sparseness and symmetry of the S-transform kernel matrix S can be exploited to perform efficient computation. As suggested by (1), the 2D S-transform is separable. Therefore, it can be achieved through 1D transforms. Hence, we shall describe only the computation of the 1D transform.

Let z be an 8-point column vector, and a vector Z be the 1D transform of z . The following steps provide a method to determine Z efficiently from z , which is also shown in Fig. 4 as a flow-graph.

$$\begin{aligned}
 m1 &= a \times z[1] \\
 m2 &= b \times z[2] - c \times z[4] + d \times z[6] - e \times z[8] \\
 m3 &= g \times z[3] - j \times z[7] \\
 m4 &= f \times z[2] + h \times z[4] - i \times z[6] + k \times z[8] \\
 m5 &= a \times z[5] \\
 m6 &= -l \times z[2] + m \times z[4] + n \times z[6] - o \times z[8] \\
 m7 &= j \times z[3] + g \times z[7] \\
 m8 &= p \times z[2] - q \times z[4] + r \times z[6] - s \times z[8] \\
 Z[1] &= m1 + m2 \\
 Z[2] &= m3 + m4 \\
 Z[3] &= m5 + m6 \\
 Z[4] &= m7 + m8 \\
 Z[5] &= m1 - m2 \\
 Z[6] &= m4 - m3 \\
 Z[7] &= m5 - m6 \\
 Z[8] &= m8 - m7
 \end{aligned}$$

The method needs 22 multiplications and 22 additions. It follows that the 2D S-transform needs 352 ($=16 \times 22$) multiplications and 352 additions, for a total of 704 operations.

The pixel domain implementation, as illustrated in Fig. 2, includes one IDCT and four HT operations. Chen's fast IDCT implementation [4], referred to herein as *the reference IDCT*, needs 256 ($=16 \times 16$) multiplications and 416 ($=16 \times 26$) additions. Each HT needs 16 ($=2 \times 8$) shifts and 64 ($=8 \times 8$) additions [5]. The four HT then need 64 shifts and 256 additions. It follows that the overall computational requirement of the pixel domain processing is 256 multiplications, 64 shifts and 672 additions, for a total of 992 operations.

Thus, the fast S-transform saves about 30% of the operations when compared to the pixel domain implementation. In addition, the S-transform can be implemented in just two stages, whereas the conventional pixel domain processing using the reference IDCT requires six stages, where the reference IDCT needs four and the HT needs two.

4 Integer Approximation of DCT-to-HT Conversion

Floating-point operations are generally more expensive to implement than integer operations. Therefore, we also provide an integer approximation of the S-transform.

We multiple S by an integer that is a power of two, and use the integer transform kernel matrix to perform the transform using an integer-arithmetic. Then, the resulting coefficients are scaled down by proper shifting. In video transcoding applications, the shifting operations can be absorbed in the quantization. Therefore, no additional operations are required to use integer arithmetic.

The larger the integer we select, the better accuracy we may achieve. In many applications, the number is limited by the microprocessor on which the transcoding is performed. We describe how to choose the number such that the computation can be performed using a 32-bit arithmetic, which is within the capability of most microprocessors.

For the case of the DCT-to-HT conversion, the DCT coefficients lie in the range of -2048 to 2047. This is a dynamic range of 4096, and needs 12 bits to represent. The maximum sum of absolute values in any row of S is 6.44, so the maximum dynamic range gain for the 2D S-transform is $6.44^2 = 41.47$, which means $\log_2(41.47) = 5.4$ extra bits. Therefore, 17.4 bits are needed to represent the final S-transform results. To be able to use the 32-bit arithmetic, the scaling factor must be smaller than the square root of $2^{32-17.4}$, i.e. 157.4. The maximum integer satisfying this condition while being a power of two is 128.

Therefore, the integer transform kernel matrix is $SI = \text{round}\{128 \times S\}$. Similar to S , SI has the form (6), but with the values a through s changed to

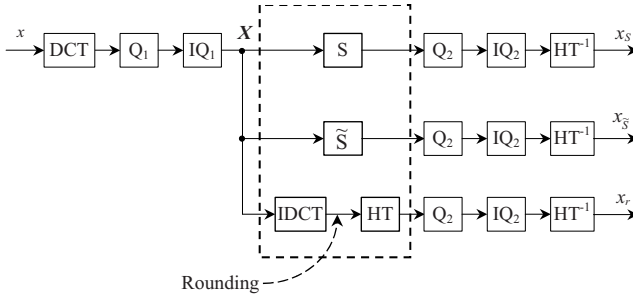


Fig. 5. Simulation settings

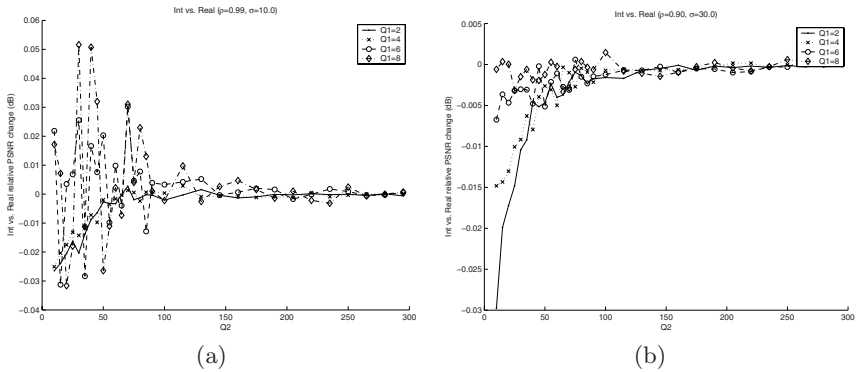


Fig. 6. Relative PSNR difference of integer vs. real S-transform

the following integers:

$$\begin{aligned}
 a &= 181, b = 164, c = 58, d = 38, e = 33, \\
 f &= 118, g = 285, h = 228, i = 111, j = 20, \\
 k &= 62, l = 14, m = 93, n = 139, o = 68, \\
 p &= 15, q = 12, r = 133, s = 253.
 \end{aligned}$$

The fast algorithm derived in Sect. 3 for the S-transform can be applied to the above transform since SI and S have the same symmetric property.

5 Simulation Results

5.1 Simulation Conditions

Fig. 5 shows the simulation setting. An 8×8 block, x , is DCT-transformed, quantized (Q_1) and inverse-quantized (IQ_1). The reconstructed coefficients, X , are sent to three processing systems. Each of them map X into the HT-domain, which are then reconstructed through quantization (Q_2), inverse-quantization

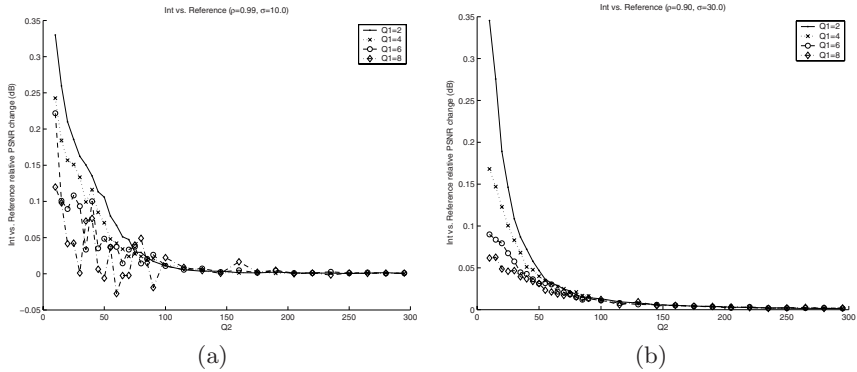


Fig. 7. Relative PSNR difference of integer S-transform vs. reference pixel domain conversion

(IQ_2), and inverse-transform (HT^{-1}). The DCT-to-HT conversion schemes used are: the real-arithmetic S-transform (S), the integer-arithmetic S-transform (\tilde{S}), and the reference IDCT-HT (IDCT followed by HT). The three reconstructed pixel-blocks are denoted as x_S , $x_{\tilde{S}}$ and x_r respectively. We use PSNR to measure the distortion between the source and the reconstructed pixel-blocks.

In the simulations, x is generated using a stationary Gaussian random process, with zero mean and standard deviation σ . If u is the distance between two pixels in x , their correlation coefficient is ρ^u . We calculate the parameters for 200 frames of the following sequences: Akiyo, Stefan, Container and Mobile&Calendar. We choose (ρ, σ) to be $(0.99, 10)$ and $(0.90, 30)$ to cover these typical sequences. We use $Q_1=2, 4, 6, 8$, and Q_2 from 10 to 295. Note that Q_2 refers to the quantization step size, not the quantization parameter (QP) in H.264/AVC bitstream. Please refer to [6] for the relationship between QP and quantization step size. We take an average of 10000 runs for each experiment.

5.2 Integer S-Transform Versus Real S-Transform

Fig. 6 shows the PSNR difference between using the integer (\tilde{S}) and the real S-transform (S). The PSNR loss resulted from the integer approximation decreases with increasing Q_2 and/or Q_1 . This is expected since the increasing quantization error makes the approximation error less significant. In addition, when Q_2 is low, the PSNR differences vary with Q_1 , and appear to be signal dependent. Interestingly, for some values of Q_1 and Q_2 , \tilde{S} achieves a PSNR gain. Nevertheless, the PSNR difference is small, with the maximum difference around 0.03dB.

In practice, it may be hardly useful to have a finer re-quantization (Q_2) than the input quantization (Q_1) since it is impossible to improve upon the quality of the input coded video. The equivalent quantization step sizes in the DCT domain and the HT domain are different due to the non-orthonormal HT. The equivalent quantization step size in the HT domain is about 5~6 times that in

the DCT domain for fine quantizations. Therefore, it can be observed that in the practical use range of Q_2 , the quality loss is even less (around 0.01dB or less).

5.3 Integer S-Transform Versus Reference IDCT-HT

Fig. 7 shows the PSNR difference between using the integer S-transform and the reference IDCT-HT. Interestingly, the integer S-transform actually outperforms the reference implementation. The inferior performance of the reference method is caused by the rounding operation after the IDCT. The rounding is the standard decoding step following IDCT in order to reconstruct pixels. In addition, the HT is an integer-transform and demands integer input. For the integer S-transform, the rounding takes place for the result HT coefficients, but not for the intermediate results. The improvement of the integer S-transform could be as much as almost 0.35dB for $Q_1=2$ and $Q_2=10$. For practical transcoding, where only coarser re-quantization may be useful, the PSNR gain is generally within 0.2dB. When Q_2 and/or Q_1 increases, the gain diminishes as the quantization error dominates the distortion.

6 Concluding Remarks

We introduced a transform domain approach for the conversion of DCT coefficients to HT coefficients. We showed that the DCT-to-HT conversion can be implemented in the transform domain by a single transformation. We derived the transformation kernel matrix and developed efficient algorithms for computing the transform. We also provided an integer approximation of the transform. Simulation results showed that the proposed transformation achieved improved PSNR performance over the conventional pixel domain implementation while requiring reduced computational complexity. Our developed transform domain DCT-to-HT conversion can be applied to the transcoding of DCT-based video such as MPEG-2 to HT-based H.264/AVC video.

References

1. ISO/IEC 13818-2: Information technology - Generic coding of moving pictures and associated audio information: Video. Edition 2 (2000)
2. Wiegand, T., Sullivan, G.J., Bjøntegaard G., Luthra A.: Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7 (2003) 560-576
3. Keesman, G., Hellinghuizen, R., Hoeksema, F., Heideman, G.: Transcoding of MPEG bitstreams. *Signal Processing: Image Communication*, vol. 8 (1996) 481-500
4. Chen, W.H., Smith, C.H., Fralick, S.C.: A Fast Computation Algorithm for The Discrete Cosine Transform. *IEEE Trans. Commun.*, vol. COM-25 (1977) 1004-1009
5. Malvar, H.S., Hallapuro, A., Karczewicz, M., Kerofsky, L.: Low-Complexity Transform and Quantization in H.264/AVC. *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7 (2003) 598-603
6. Richardson, I.E.G.: H.264 / MPEG-4 Part 10 Tutorials: H.264 Transform and Quantization. Available online at <http://www.vcodex.com/h264.html>.

An Adaptive Hybrid Mode Decision Scheme for H.264/AVC Video

Feng Huang¹, Jenq-Neng Hwang², and Yuzhuo Zhong¹

¹ Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
`hfbee@263.net`, `zyz-dcs@tsinghua.edu.cn`

² Department of Electrical Engineering, Box 352500,
University of Washington, Seattle, WA 98195, USA
`hwang@ee.washington.edu`

Abstract. In this paper, a hybrid rate-distortion (RD)-based mode switching decision scheme is proposed for robust transmission of H.264/MPEG4-AVC video over unreliable networks. Similar to the recursive optimal pixel estimate (ROPE) algorithm, the proposed encoder recursively estimates the overall distortion of decoder frame reconstruction due to quantization, error propagation, and error concealment. The estimation is then used for switching between intra-coding and inter-coding modes per macro-block. Different from ROPE, our scheme assumes the worst case of the previous frame to make RD optimal mode switching decision for the current frame. To avoid the reduction of coding efficiency, the worst scenario is selectively applied according to the motion activity of the previous frame, which is measured by the potential concealment error. By adopting the notion of concealment candidate image (CCI), we successfully implement the Hybrid-ROPE (HROPE), which finds a good tradeoff between error resilience and coding efficiency. Thus, the HROPE algorithm yields consistent and significant gains over the competing methods. The experimental results under different simulation conditions show that our mode decision scheme has robust error resilient ability and good adaptability. With the advanced feature of H.264, flexible MB ordering (FMO), our scheme can be combined with more existing error concealment techniques. It also attributes to FMO that the ROPE-styled framework in our scheme provides the flexibility of the packet sizes in terms of the maximum transfer unit (MTU) size. Moreover, the hybrid mode decision scheme has low computational complexity and does not require a feedback channel.

1 Introduction

The internet is ill-suited for the transmission of time-critical data such as interactive video. Packets can be lost or dropped when intermediate links or routers become congested due to excessive traffic. Resilience to packet loss is a critical requirement in predictive video coding for transmission over packet-switched networks, since the prediction loop propagates errors and causes substantial

degradation in video quality. The mismatch of reference frame between encoder and decoder is the major factor that causes degradation of video quality. In order to avoid error propagation over a long period, the encoder can use an intra-coding refreshment scheme to refresh the decoding process and stop (spatial and temporal) the error propagation. However, too many intra-coded macro-blocks (MBs) will certainly decrease the coding efficiency. Thus, the tradeoff between compression efficiency and error resilience is very important and has been widely addressed.

An adaptive Intra Refreshment (IR) scheme implements mode selection within a rate-distortion (RD) optimization framework so as to directly optimize the performance. The rate-distortion optimized MB mode selection algorithms use a Lagrangian cost function that linearly combines terms of “rate” and “distortion”. Moreover, an estimate of the expected distortion caused by transmission errors and losses can also be taken into account in the cost function. A number of distortion estimating algorithms [1][2] have been proposed, and the loss-aware rate-distortion-optimized (LA-RDO) MB mode selection algorithm was adopted by H.264 for reference implementation [3]. The LA-RDO method has significantly better performance than non-adaptive and other adaptive algorithms. However, its computational complexity is typically multifold, which limits the use in practical implementations. The recursive optimal per-pixel estimate (ROPE) algorithm [4] estimates all the distortion of the decoder frame reconstruction due to quantization, error propagation and error concealment. The estimation is then used for switching between intra-coding and inter-coding modes per macro-block in a RD-based framework. ROPE combines the previous frame due to quantization and the previous frame due to concealment to produce the weighted frame. However, since it is reasonable to assume the packet-loss rate is less than 20% in video transmission, the weighted frame is just slightly different from the reconstructed frame due to quantization. Thus, the weighted frame does not effectively reflect the degradation of the previous frame caused by channel error. This affects the accuracy of the estimation and the intra/inter mode decision of the current frame. Based on the statistical model of error propagation and the introduction of the concealment candidate image (CCI) [5], this paper proposes an improved RD optimal mode switching decision scheme, which has the following improvements:

- 1: The CCI of the previous frame is adopted to evaluate the worst deterioration of the previous frame. The encoder will select more intra MB mode for the current frame to stop error propagation by taking into account potential packet loss.
- 2: To avoid the reduction of coding efficiency due to selecting more intra MBs, we compare potential concealment error of the previous frame with a typical threshold. This helps to decide whether the worst deterioration of the previous frame or the modest one should be used for the intra/inter mode decision of the current frame. Since the potential concealment error is used as a physical measure to reflect the time-varying feature of video content, the number of intra MBs in the current frame reflects the motion activity of the previous frame.

A more advanced feature, called the flexible MB ordering (FMO) [6], has also adopted in H.264/AVC. FMO permits the specification of different patterns for the mapping of the MBs to slices including checker-board-like patterns, sub-pictures within a picture, or a dispersed mapping of MBs to slices. We choose the checker-board-like FMO pattern for our proposed RD-based framework due to the followings:

- 1: Our scheme can be applied with more existing error concealment schemes with FMO than without FMO since the samples of a missing slice are surrounded by many samples of correctly decoded slices in terms of the checker-board-like pattern.
- 2: ROPE algorithm [4] assumes that all the MBs in a particular slice come from the same row, which limits the flexibility of coded slice sizes. It is advisable to keep coded slice sizes as close to, but never bigger than, the MTU size [7]. Nevertheless, the MTU size for wireline IP links is quite different from the one in a wireless environment. Moreover, the end-to-end MTU size of a transmission path between two IP nodes may change dynamically during a connection. FMO provides the flexibility of the coded slice sizes to deal with the diversity of the MTU size in the ROPE-styled framework of our scheme.

This paper is organized as follows. In Section 2, we analyze distortion and error propagation problems and the use of FMO patterns in H.264 video coding. In Section 3, we derive an effective mode-switching algorithm based on RD optimal framework, and use a hybrid recursive algorithm that computes, at the encoder, the optimal estimate of the overall distortion of decoder reconstruction at the pixel level precision. We present results to demonstrate the performance of the method in Section 4. The superiority of the proposed approach over other state-of-the-art mode switching techniques is demonstrated by simulation. Conclusions are drawn in Section 5.

2 Distortion and Error Propagation with Error-Prone Channels

2.1 Preliminaries

Flexible MB ordering (FMO) allows assigning MBs to slices in an order other than the scan order. To do so, each MB is statically assigned to a slice group using a MB allocation map. In Fig. 1, three rows of MBs of the picture are allocated either slice group 0 or 1, depicted in grey and white, respectively, in a checker-board fashion.

In our coding system, we form a slice from a few rows of MBs in the same slice group, and assume that each slice is carried in a separate packet. In this setting, the loss rate of a MB equals the packet loss rate p . We assume that the packet loss rate p is available at the encoder. This can be either specified as part of the initial negotiation, or adaptively calculated from information provided by the transmission protocol, such as RTCP [8].



Fig. 1. Three rows of a QCIF picture and two slice groups

Let f_n^i denote the original value of pixel i in frame n , and let $f_n^{\wedge i}$ denote its *encoder* reconstruction. The constructed value at the *decoder*, possibly after error concealment, is denoted by $f_n^{\sim i}$. Recall that for the encoder, $f_n^{\sim i}$ is a random variable. Using the SAD (the sum of absolute difference) as distortion metric, the overall expected distortion for this pixel is

$$D_n = E\{|f_n^i - f_n^{\sim i}|\} = D1_n + D2_n, \tag{1}$$

where $D1_n$ is the distortion calculated assuming the corresponding MB of the current frame n , is correctly received, and $D2_n$ is the concealment distortion assuming the corresponding corrupted MB. As follows, we consider two cases depending on whether the pixel belongs to an intra-coded MB or an inter-coded MB.

2.2 Distortion in Intra-coded MB

Let us first assume that an intra MB is received correctly at the decoder. We thus have $f_n^{\sim i} = f_n^{\wedge i}$, and the probability of this event is $1-p$. Then we derive $D1_n$ for the intra-coded MB

$$D1_n(i) = (1 - p)|f_n^i - f_n^{\wedge i}|. \tag{2}$$

If the MB is lost, the decoder first checks if its left MB and its right MB have been received correctly. Recall that the left MB and the right MB belong to the same slice according to the check-board fashion. If the slice is available, the estimated motion vector of the lost MB is calculated as the median motion vector of its left MB's and its right MB's. The estimate is used to associate pixel i in the current frame with pixel k in the previous frame. We thus have $f_n^{\sim i} = E\{f_{n-1}^{\sim k}\}$, and the probability of this event is $p(1-p)$. On the other hand, if the left one and the right one of the lost MB are lost as well, we set the estimated motion vector to zero. Therefore, we have $f_n^{\sim i} = E\{f_{n-1}^{\sim i}\}$ with probability p^2 . Combining the two cases, we derive $D2_n$ for the intra-coded MB

$$D2_n(i) = p(1 - p)|E\{f_{n-1}^{\sim k}\} - f_n^i| + p^2|E\{f_{n-1}^{\sim i}\} - f_n^i|. \tag{3}$$

2.3 Distortion in Inter-coded MB

In inter-mode, the encoder predicts the current pixel i from the pixel j in the previous frame by the true motion vector of the MB. Thus, the encoder prediction of this pixel is $f_{n-1}^{\wedge j}$. The motion prediction residue is then compressed and we

denote the quantized residue by $e_n^{\wedge i}$. Thus we have $e_n^{\wedge i} = f_n^{\wedge i} - f_{n-1}^{\wedge j}$. If the current MB is correctly received, the decoder has access to both $e_n^{\wedge i}$ and the motion vector. We thus have $f_n^{\sim i} = e_n^{\wedge i} + E\{f_{n-1}^{\sim j}\}$. Then we derive $D1_n$ for the inter-coded MB

$$D1_n(i) = (1 - p)|f_n^i - (e_n^{\wedge i} + E\{f_{n-1}^{\sim j}\})| . \tag{4}$$

If the packet containing the inter-coded MB is lost, the decoder performs error concealment in a manner identical to that of an intra-coded MB. So the concealment distortion $D2_n$ is identical for both intra-mode and inter-mode. Note that these recursions are performed at the encoder in order to calculate the expected distortion at the decoder. The encoder can exploit this result directly for mode switching.

3 RD-Based Mode Switching Algorithm for H.264

RD optimization algorithm is a useful tool for video compression in error-free channels [9]. H.264 adopts it to find the best motion vector, the best reference frame, the best intra prediction mode, and to select the best MB mode. H.264 supports not only the multiple inter-modes (16x16, 16x8, 8x16, 8x8, 8x4, 4x8, 4x4) with different block types, but also skip-mode and intra-modes (4x4, 16x16). The high complexity mode of H.264, however, doesn't consider the distortion caused by channel error. A loss-aware RD-optimized MB mode selection algorithm (LA-RDO), has been selected into H.264 reference implementation [3]. In the encoder, K copies of the random channel behavior and the hypothetical decoder are manipulated. The expected distortion at the decoder can be estimated rather accurately if K is chosen large enough. However, the added complexity in the encoder is obviously at least K times the decoder complexity. Therefore, LA-RDO is limited in practical implementations.

We use the same RD optimal framework as that in H.264. Considering $D2_n$ is identical regardless of the MB mode, we simplify the computation of the cost as

$$\min_{mode} (J_{MB}) = \min_{mode} (D1_{MB} + \lambda R_{MB}) , \tag{5}$$

where the distortion of the MB is the sum of the distortion contributions of the individual pixels

$$D1_{MB} = \sum_{i \in MB} D1_n(i) , \tag{6}$$

Note that R_{MB} denotes the bit rate, and λ is Lagrange multiplier. The optimal encoding mode for each MB can thus be determined by (5).

3.1 Recursive Algorithm Based on Concealment Candidate Image (CCI)

After the previous frame $n-1$ is encoded, we calculate $E\{f_{n-1}^{\sim i}\}$ of intra-coded MB as

$$E\{f_{n-1}^{\sim i}\} = (1 - p)f_{n-1}^{\wedge i} + p(1 - p)E\{f_{n-2}^{\sim k}\} + p^2E\{f_{n-2}^{\sim i}\} , \tag{7}$$

for the mode decision of frame n . Similarly, we can calculate $E\{f_{n-1}^{\sim i}\}$ of inter-coded MB as

$$E\{f_{n-1}^{\sim i}\} = (1-p)(e_n^{\wedge i} + E\{f_{n-2}^{\sim j}\}) + p(1-p)E\{f_{n-2}^{\sim k}\} + p^2E\{f_{n-2}^{\sim i}\}, \quad (8)$$

However, since it is reasonable to assume that p is less than 20% in video packet transmission, the weighted construction $E\{f_{n-1}^{\sim i}\}$ is slightly different from the encoder reconstruction $f_{n-1}^{\sim i}$. Thus, $E\{f_{n-1}^{\sim i}\}$ does not effectively reflect the quality degradation to the frame $n-1$ due to channel error, especially in minor or high packet-loss environments. This affects the accuracy of the estimation and the intra/inter mode decision of the frame n .

It is advisable to displace $E\{f_{n-1}^{\sim i}\}$ in the RD-based mode decision of frame n . Thus, we adopt the concealment candidate image (CCI) [5] of frame $n-1$. More specifically, each pixel in CCI is computed as

$$C\{f_{n-1}^{\sim i}\} = (1-p)E\{f_{n-2}^{\sim k}\} + pE\{f_{n-2}^{\sim i}\}, \quad (9)$$

whether the pixel is in intra-coded MB or not. In essence, each pixel value in CCI- $n-1$ is the same as the concealed one due to channel error.

Since $C\{f_{n-1}^{\sim i}\}$ reflects the worst deterioration of frame $n-1$, we refine $D1_n$ of inter-coded MB (4) as

$$D1_n(i) = (1-p)|f_n^i - (e_n^{\wedge i} + C\{f_{n-1}^{\sim j}\})|, \quad (10)$$

where $D1_n$ of intra-coded MB follows Eq. (2). As a result, we propose the recursive algorithm as follows.

1. For intra mode, calculate $D1_{MB}$ for the MB of frame n via Eqs. (2), (6).
2. For inter mode, calculate $D1_{MB}$ for the MB of frame n via Eqs. (10), (6).
3. Select the best mode for the MB of frame n via Eq. (5).
4. If any MB of frame n has not been encoded via Step 1-3, then go to Step 1.
5. Update $E\{f_n^{\sim i}\}$, $C\{f_n^{\sim i}\}$ for each pixel of frame n . Increase n , and then go to Step 1.

3.2 Hybrid Recursive Algorithm

Since $D1_n$ usually increases after the refinement from Eqs. (4) to (10), the encoder will select more intra MB mode for frame n to stop error propagation. To avoid the reduction of coding efficiency due to more intra MBs, we apply the concealment error image (CEI) [5] for frame $n-1$, and compute the potential mean concealment error (MCE) of frame $n-1$ as

$$MCE_{n-1} = \sum_{i \in CCI-n-1} |f_{n-1}^{\wedge i} - C\{f_{n-1}^{\sim i}\}|/(XY), \quad (11)$$

where X, Y is the width and height of CCI. If MCE_{n-1} is larger than a typical threshold T , we compute $D1_n$ via Eq. (10); otherwise, we compute $D1_n$ via Eq. (4). This helps to make the number of intra MBs in frame n reflect the motion activity of frame $n-1$ without sacrificing too much coding efficiency. Therefore, we propose a hybrid recursive algorithm as follows.

1. For intra mode, calculate $D1_{MB}$ for the MB of frame n via Eqs. (2), (6).
2. For inter mode, if $MCE_{n-1} > T$ and $D1_{n-1}$ is computed via Eq. (4), calculate $D1_{MB}$ for the MB of frame n via Eqs. (10), (6). Otherwise, calculate $D1_{MB}$ for the MB of frame n via Eqs. (4), (6).
3. Select the best mode for the MB of frame n via Eq. (5).
4. If any MB of frame n has not been encoded via Step 1-3, then go to Step 1.
5. Update $E\{f_n^{\sim i}\}$, $C\{f_n^{\sim i}\}$, for each pixel of frame n . Increase n , and then go to Step 1.

4 Simulation Results

We implemented the two proposed recursive algorithms, as presented in 3.1 and 3.2, for mode decision by appropriately modifying the JVT Software Version M7.3. CIF sequences are encoded at the specified quantizing parameter by the H.264 baseline encoder. We assume the RTP payload format for packetizing the H.264 video stream [10]. Each packet contains only one slice and one slice contains several rows of MBs in the same slice group. We denote the number of MBs in one packet as N_{MB} . A random packet loss generator is used to drop packets at a specified loss rate p .

We compare the performance of five RD frameworks for MB mode decisions. They are listed as follows.

1. The encoder doesn't consider the channel error in the H.264 draft (RDO) [3].
2. The encoder simulates the error prone channel by multiple decoders [3] and makes the mode decision. This is the loss-aware RD optimized mode selection algorithm (LA-RDO). The number of hypothetical decoders, K , is set to 30 in all simulations.
3. The H.264 encoder adopts the recursive optimal per-pixel estimate (ROPE) algorithm for mode switching.
4. Our recursive optimal per-pixel estimate algorithm based on CCI (ROPE-CCI).
5. Our hybrid recursive optimal per-pixel estimate algorithm (HROPE). T is set to 6 in all simulations.

The peak signal-to-noise ratio (PSNR) of the decoder luminance reconstruction is computed for each frame and averaged over the whole sequence.

Fig. 2 shows the PSNR versus the bit rate of the five methods for the "Foreman" sequence with packet loss rate of 15% when one packet contains six rows of MBs in the same slice group. The result supports the claim that the proposed HROPE algorithm yields consistent and significant gains over the three methods: ROPE, LA-RDO and RDO. However, when compared with HROPE, the proposed ROPE-CCI shows less improvement over the three methods. As stated before, ROPE-CCI assumes the worst deterioration of the previous frame so that the coding efficiency is reduced. HROPE adaptively assumes the worst deterioration of the previous frame according to its motion activity.

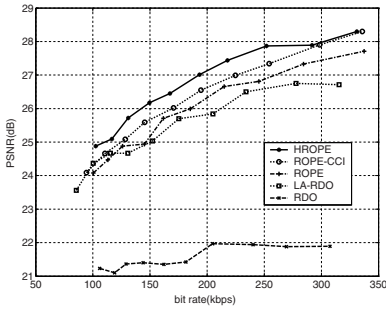


Fig. 2. PSNR versus bit rate of five RD methods: HROPE (proposed), ROPE-CCI (proposed), ROPE [4], LA-RDO[3], RDO[3]. Foreman CIF sequence, $N_{MB} = 66$, $p = 15\%$, $f = 15f/s$

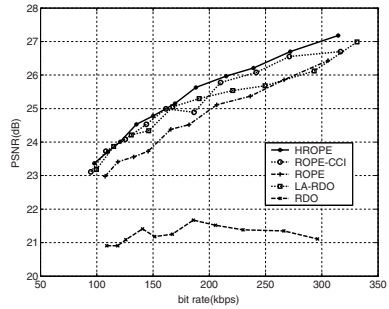


Fig. 3. PSNR versus bit rate of five RD methods: HROPE (proposed), ROPE-CCI (proposed), ROPE [4], LA-RDO[3], RDO[3]. Foreman CIF sequence, $N_{MB} = 22$, $p = 15\%$, $f = 15f/s$

Our proposed algorithms also achieve performance gains over the three competing methods when $N_{MB} = 22$. That means the coded packet size is changed to deal with a different MTU size. Fig. 3 shows the PSNR versus the bit rate of the five methods for the Foreman sequence when $N_{MB} = 22$ and $p = 15\%$.

The relative performance of the five methods depends on the simulation conditions. The advantage of our proposed algorithms decreases and even becomes counterproductive in low packet-loss environments or for sequences of low motion. Fig. 4 shows the PSNR versus the bit rate of the five methods for the foreman sequence when $N_{MB} = 66$ and $p = 3\%$. Fig. 5 shows the PSNR versus the bit rate of the five methods for the Akyio sequence when $N_{MB} = 66$ and $p = 15\%$. In the two simulations, the performance of ROPE is close to the one of HROPE. Moreover, ROPE outperforms our proposed ROPE-CCI. Nevertheless, HROPE finds a good tradeoff between error resilience and coding efficiency and is always not inferior to the other RD methods besides ROPE-CCI.

It is possible to encounter mismatch between the packet loss rate assumed at the encoder and the actual loss rate at the decoder. Fig. 6 describes the performance when the encoder assumes $p = 5\%$ while the actual packet loss rates are 1%, 3%, 5%, 7%, or 9%. Both the proposed HROPE and ROPE-CCI show good robustness to mismatch. Moreover, HROPE increases its performance gains over ROPE, LA-RDO and RDO as the actual packet loss rate grows.

5 Conclusions

In this paper, a hybrid RD-based mode switching decision scheme is proposed for robust transmission of H.264/MPEG4-AVC video over unreliable networks. Similar to the recursive optimal pixel estimate (ROPE) algorithm, we intend to make our H.264 error-resilient encoder ‘farseeing’ enough to reconstruct the distortion at the decoder.

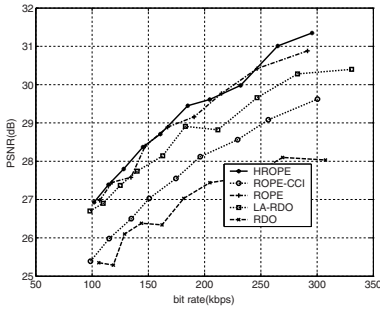


Fig. 4. PSNR versus bit rate of five RD methods: HROPE (proposed), ROPE-CCI (proposed), ROPE [4], LA-RDO[3], RDO[3]. Foreman CIF sequence, $N_{MB} = 66$, $p = 3\%$, $f = 15f/s$

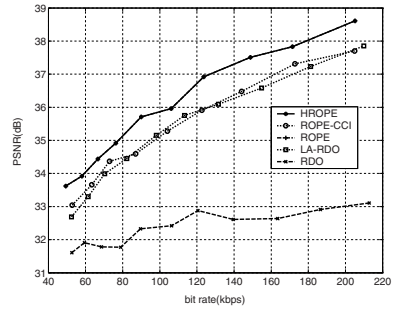


Fig. 5. PSNR versus bit rate of five RD methods: HROPE (proposed), ROPE-CCI (proposed), ROPE [4], LA-RDO[3], RDO[3]. Akyio CIF sequence, $N_{MB} = 66$, $p = 15\%$, $f = 15f/s$

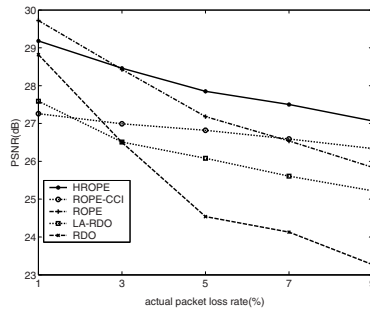


Fig. 6. PSNR versus actual packet loss rate at the decoder. Foreman CIF sequence, 5% assumed, $bitrate = 150kbps$, $N_{MB} = 66$, $f = 15f/s$

We highlight two features of our scheme to distinguish from ROPE. Firstly, it assumes the worst case of the previous frame to make RD optimal mode switching decision for the current frame. Secondly, the potential concealment error is used to measure the motion activity of the previous frame. Then the first feature is selectively applied according to the result of such measurement. As the key algorithm of our scheme, HROPE, extended from another proposed algorithm ROPE-CCI, finds a good tradeoff between error resilience and coding efficiency. The experimental results under different simulation conditions show that our mode decision scheme has robust error resilient ability and good adaptability. The HROPE algorithm yields consistent and significant gains over the competing methods. This advantage decreases for sequences of low motion or for low packet-loss environments. However, HROPE achieves the best performance among the RD-based mode switching methods.

We also integrate the flexible MB ordering (FMO), an advanced feature of H.264, into our scheme. With FMO our scheme can be applied with more existing error concealment schemes provided that the same error concealment is used at both encoder and decoder. It also attributes to FMO that the ROPE-styled frameworks provide the flexibility of the packet sizes. Thus, our scheme can achieve substantial gains in both wired and wireless packet-lossy networks, where the MTU sizes are quite different. Moreover, the hybrid mode decision scheme has low computational complexity and does not require a feedback channel.

References

1. E. Steinbach, N. Farber, and B. Girod, "Standard compatible extension of H.263 for robust video transmission in mobile environments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 872-881, Dec. 1997.
2. A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 23-50, Nov. 1998.
3. ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC Draft Text, Joint Video Team document JVT-E146d37, Oct. 2002.
4. R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal intra/inter mode switching for packet loss resilience," *IEEE Journal of Selected Areas in Communications*, vol. 18, no. 6, pp. 966-976, June 2000.
5. W. J. Chen, J. N. Hwang, "The CBERC: A Content-Based Error-Resilient Coding Technique for Packet Video Communications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 8, August 2001.
6. S. Wenger and M. Horowitz, "Flexible MB ordering-A new error resilience tool for IP-based video," *presented at the IWDC 2002*, Capri, Italy, Sept. 2002.
7. S. Wenger, "H.264/AVC Over IP," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, July 2003.
8. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," *RFC'89*, Jan. 1996.
9. G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 74-90, Nov. 1998.
10. S. Wenger, T. Stockhammer, and M. M. Hannuksela, "RTP payload format for H.264 video," *in Internet Draft, Work in Progress*, Mar. 2003.

Workload Characterization of the H.264/AVC Decoder

Tse-Tsung Shih, Chia-Lin Yang, and Yi-Shin Tung

Department of Computer Science and Information Engineering
National Taiwan University
{r91099, yangc, r5526029}@csie.ntu.edu.tw

Abstract. Multimedia applications have become important workloads for modern computer systems. The latest video coding standard, H.264/AVC, adds new features to improve the coding efficiency and visual quality at the cost of higher implementation complexity. The increasing computation and storage requirements pose challenges to achieve real-time video playback on general-purpose processors (GPPs). In this paper, we aim at identifying the performance bottleneck of running a software implementation of the H.264/AVC decoder on GPPs. We adopt a simulation-based approach to perform detailed workload characterization.

1 Introduction

Multimedia applications have become important workloads for modern computer systems. The latest video coding standard, H.264/AVC [1,2], seeks to provide high coding efficiency at very low bit rates. H.264 follows the same block-based motion compensation and transform coding framework as the existing standards, such as MPEG2. However, it adds a number of new features and functionalities to improve the coding efficiency. These added features also result in additional complexity in encoding and decoding. The increasing computation and storage requirements of H.264/AVC pose challenges to achieve real-time video decoding on GPPs (General Purpose Processors). In this paper, we study and analyze the performance of a software implementation of H.264/AVC decoder on GPPs. Through this study, we can find out the performance bottleneck of running the H.264/AVC decoder on a modern GPP and what features of the H.264/AVC decoder cause the bottlenecks.

There are two objectives of performing workload characterization on the H.264 decoder. First, understanding the characteristics of the H.264/AVC decoder allows us to tune processor architecture to suit particular features in program. Second, it can guide software developers for the H.264/AVC decoder to tune their programs for performance. In this study, we focus on the following program characteristics of the H.264/AVC decoder:

(1) The available instruction level parallelism

Modern microprocessors rely on aggressive out-of-order execution techniques to exploit instruction level parallelisms (ILP) present in applications. The intrinsic available ILP in an application has direct impact on its performance on a superscalar processor.

(2) Program locality

As the performance gap between processors and memory continues to grow, techniques to reduce the effect of this disparity are essential to build a high performance computer system. The use of caches between the CPU and main memory is recognized as an effective method to bridge this gap. If programs exhibit good temporal and spatial locality, the majority of memory requests can be satisfied by caches without having to access main memory. Cache misses inhibit the ability of a superscalar processor to exploit available instruction level parallelisms. Therefore, program locality is critical to performance.

(3) Predictability of control structures

Modern processors employ speculative execution techniques to reduce the CPU stall time due to branch instructions. They rely on branch prediction mechanisms to guess the direction of a branch instruction (taken or fall-through), and allow execution to proceed on the predicted direction. Programs with unpredictable control flows cause significant CPU stall time due to branch misprediction.

To understand what features of the H.264/AVC decoder cause the performance bottleneck, in addition to analyze the aforementioned characteristics for the whole application, we also look into the behavior of individual kernels. Furthermore, we investigate what application features (sequence content, resolution, bitrate) and new coding tools (multi-ref frames, CABAC) have direct impact on its performance on a superscalar processor.

In this study, we adopt the simulation-based approach and evaluate the characteristic of H.264 comprehensively. Previous works that study the H.264/AVC decoder either perform algorithmic complexity analysis [7] or rely on on-chip monitoring counter to gather performance profile [8]. To our knowledge, this is the first work that use the simulation-based method to characterize the H.264/AVC decoder. Although the simulation-based approach is more time-consuming. It allows us to explore the design space thoroughly and evaluate different architectural enhancements. Below we summarize the important findings in this study:

- (1) The H.264/AVC decoder does present significant instruction level parallelism. Assuming perfect memory system, branch prediction and unlimited resources, we can see performance scales as the issue width increases. The performance scaling levels off until the issue width is up to 32.
- (2) The H.264/AVC decoder is computation-bound not memory-bound. A commonly held belief is that multimedia applications often have large data sets and seem to have little data reuse. However, the simulation results show that

memory subsystem only contributes a small portion of CPU stall time. The reason is that there exist significant data reuse in macroblock level data, and that can be captured in a 16-K data cache.

- (3) The H.264/AVC decoder show high branch mispredicted ratio assuming a two-level branch predictor. Variable length coding and deblocking filter are two kernels that show worst branch behavior. The high branch misprediction rate is due to nested loops and content dependent branches. Loop unrolling and special instructions, such as absolute value calculation, can reduce CPU stall time due to mispredicted branches significantly.
- (4) On the application side, we find that video contents with low motion and smaller resolution increase the inter frame prediction opportunity thereby increasing cache miss rates. That is because the data cache only captures data reuse at the macroblock level. Higher bitrates increase execution time of variable length decoding thereby increasing CPU stall time due to branch misprediction. New added multi-ref frame feature does not have direct impact on cache performance since inter frame reuse cannot be captured by the data cache. CABAC has lower control flow predictability than CAVLC due to bit-wise access to bitstream.

The remainder of this paper is organized as follows. Section 2 gives an overview of H.264/AVC video coding standard. Section 3 describes our experimental methodology. Section 4 characterizes the behavior of the H.264/AVC decoder on a superscalar processor. Finally, conclusions are given Section 5.

2 H.264/AVC Overview

Since finalization of H.263 in 1998, a long-term project H.26L was enforced by ITU to collect the new or improved coding tools that advance video coding. In 2001, based on the result of H.26L, ISO MPEG and ITU VCEG co-founded a new working group called Joint Video Team (JVT) to evolve the draft to a new video coding standard. After the two-years effort, the new standard is finalized in 2003. It is entitled “Advanced Video Coding” (AVC) and is published jointly as MPEG-4 part-10 and ITU-T Recommendation H.264 [1, 2].

Although H.264/AVC keeps the basic coding structure unchanged, the new standard adopts a lot of elaborations, which can improve coding performance with amending only the marginal computation cost. These major changes include (1)More flexibility and generalization on block prediction. (2)Multiple tables for context adaptive VLC (CAVLC) coding and the context adaptive binary arithmetic coding (CABAC) [3]. (3)Small size block transform [4] and built-in deblocking filter [5]. The aforementioned new features influence the codec implementation majorly in the following aspects:

- (1) Providing too many different coding options in logical will introduce many inevitable branches.
- (2) As the basic coding block becomes smaller, i.e. 4×4 , the degree of parallel processing capability also decreases.

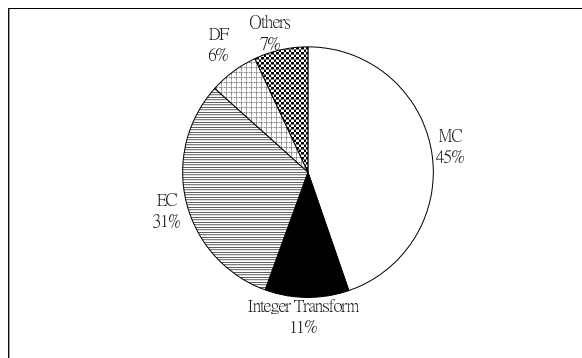


Fig. 1. Important kernel time breakdown in H.264/AVC

- (3) Adaptive deblocking requires pixel-wise processing and threshold comparisons. This process is hard to be realized efficiently or in a parallel way.
- (4) Although binary arithmetic coding is adopted to reduce the necessary complexity, CABAC is still more computation intensive than CAVLC. Bitwise access to input bitstream and several normalizations are necessary in this stage.

The execution break down of the H.264/AVC decoder is shown in Fig. 1. The H.264/AVC contains 5 major kernels: motion compensation (MC), inverse integer transform, entropy coding (EC) and deblocking filtering (DF). Note that we adopt CABAC for entropy coding to obtain the data. Motion compensation is the most time-consuming module because H.264 uses a quarter resolution. Interpolating pixels at fractional positions requires intensive computations in H.264 decoder. Entropy coding also consumes a significant percentage of overall execution time. As mentioned before, the new coding method, CABAC, adopted in H.264 increases computation complexity. Inverse integer transform does a 4×4 matrix multiplication. Deblocking filter compares the pixels at block edge and smooths them if necessary.

3 Experimental Methodology

We use the SimpleScalar toolset[9] to perform this study. The SimpleScalar is a cycle-accurate architecture simulator which models a standard 5 pipeline stage superscalar processor. It supports speculative execution using 2-level branch predictor. The software implementation of the H.264/AVC decoder used in this study is the reference software Joint Model (JM73) developed by the HHI[6]. The test sequences used in this study are summarized in Table 1. This test suite contains sequences with various motion contents. For example, the container sequence is a still camera scene with slow foreground and local motion, while the foreman sequence contains fast foreground motion. For each sequence, we also test different resolutions, QCIF (172×144), CIF (352×288) and 4CIF (720×480).

Table 1. The properties and characteristics of the selected test sequences used throughout this paper

Sequences	Resolutions	No of Frames	Characteristics
Container Ship	QCIF/CIF	300/300	Still camera scene with slow foreground and local motion.
Flower Garden	4CIF	100	Fast camera panning with foreground and background objects of different viewing distances; high spatial and texture pattern inside.
Foreman	QCIF/CIF	300/300	Fast foreground motion during the first half and fast camera panning in the second half.
Football	4CIF	100	Multiple foreground objects with irregular large motions.
Mobile&Calendar	QCIF/CIF/4CIF	300/300/100	Slow panning and zooming with complex motion; high spatial and color detail inside.

4 Workload Characterization

In this section, we present detailed analysis on the performance of the H.264/AVC decoder on a superscalar processor. We first quantify the available instruction level parallelism, and then investigate the causes for the CPU stall time. In particular, we look into the memory and control flow behaviors of the H.264/AVC decoder in details.

4.1 Instruction Level Parallelism Analysis

To quantify the available instruction level parallelism, we measure the IPC (Instruction per Cycle) varying the issue width from 4 to 128 assuming perfect memory system, branch prediction and unlimited resources. The results are shown in Fig.2 (i.e., the solid line). We can see that the IPC increases as the issue width increases, and reaches plateau when the issue width is 32. This indicates that there does exist significant instruction level parallelism in the H.264/AVC decoder. With IPC equal to 4, a 2.7 Ghz processor can actually decode a 4CIF, 11.3Mbit/s sequence in real time (i.e., 30 frames per second).¹

However, a superscalar processor is often not able to fully exploit the available ILP due to cache misses, mispredicted branches and resource constraints. In Fig. 2, we also plot the IPC on realistic processor model. For a 4-issue processor, we assume a 32k direct-map instruction cache and 16k-2way data cache, 2-level branch predictor and 32 fetch-queue, 64 reorder buffer, 32 load-store queue. We scale the resources as the issue width increases. We can see that a realistic 4-issue processor can only achieve IPC equal to 2.7. Therefore, to play 4CIF, 11.3Mbit/s sequence in real time, we will need a 4Ghz processor. The IPC of an 8-issue processor is close 4, which is almost half of the IPC of an ideal processor. Increasing issue width beyond 8 does not improve performance further. The causes of the wasted CPU time can be classified into 4 types: instruction cache misses, data cache misses, mispredicted branches and limited resources. In Fig. 3, we break down the CPU stall time of a 4-issue processor into these 4 types. In contrast to commonly held belief that multimedia applications have poor cache performance due to their streaming access patterns, we find that the memory

¹ Time = Instruction Count \times CPI \times Clock rate. We run 100 frames and get execution cycles from simulator, then we can estimate the CPU frequency that needed.

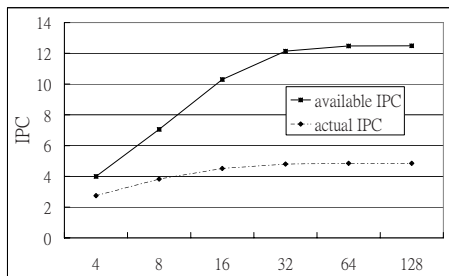


Fig. 2. Analysis of available instruction level parallelism in H.264/AVC

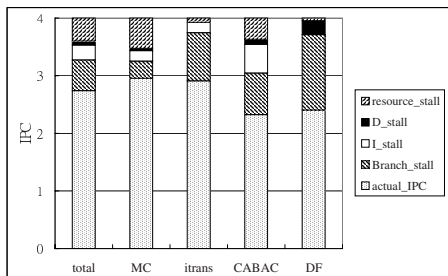


Fig. 3. IPC Breakdown by 4-way issuing simulation

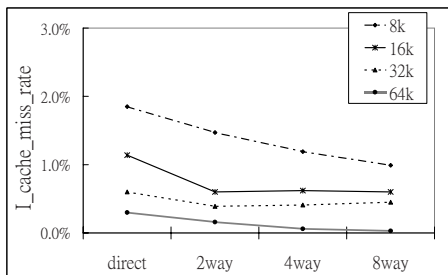


Fig. 4. Instruction cache miss rate

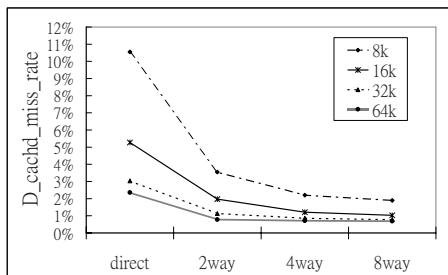


Fig. 5. Data cache miss rate

subsystem actually contributes less than 3% of the CPU stall time. Mispredicted branches appear to be the major sources of wasted CPU time. Next, we analyze the memory and control flow behaviors of the H.264/AVC in details to provide answers for the observed phenomenon.

4.2 Memory System Characteristic

To characterize the memory system behavior of the H.264/AVC decoder, we first identify cache architectures that are suitable for the H.264/AVC decoder. Fig. 4 and Fig. 5 show the design space exploration of the instruction and data cache, respectively. The block size is fixed at 32 bytes for both caches. For the instruction cache, a 32K, direct-mapped cache is able to achieve miss rates less than 1%. As for the data cache, we find that a 16K, two-way associative cache has 2% miss rate. Doubling the cache size (32K) reduces the cache miss rate to 1%. The instruction cache miss rate is lower than the data cache, however, in Fig. 3, we see more CPU stall time due to instruction cache misses. That is because memory access latency due to data cache misses can be hidden through out-of-order execution.

Although the H.264/AVC decoder presents streaming access patterns, however, there is a significant amount of data reuse in the macroblock level. H.264 partitions the input data into many smaller macroblocks, which are the basic coding units. The size of a macroblock is 16×16 pixels. The macroblock data have

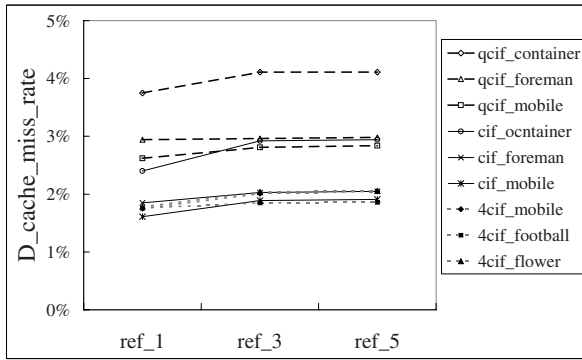


Fig. 6. Data cache misses per macroblock with 1, 3 and 5 reference frames, respectively

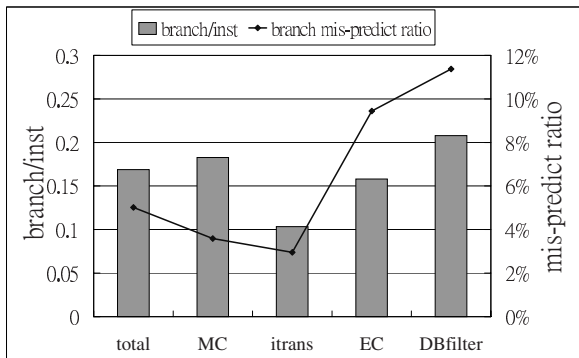


Fig. 7. Branch per instructions and misprediction rate in each kernel of H.264/AVC motion compensation

high temporal locality. During the deblocking filtering and intra-motion compensation process, the whole line of macroblocks above them are referenced. One line of macroblock data for a 4CIF sequence is about 16K bytes ($w/16 \times 16 \times 16 \times 1.5$).

Fig. 6. shows the effect of multiple reference frames, sequence resolution and sequence contents on cache performance. According to the experimental results, these three application attributes have only little effect on the data cache miss ratio.

4.3 Control Flow Behavior

The IPC breakdown analysis shown in Section reveals that CPU stalls due to mispredicted branches are the major performance bottleneck. In this section, we evaluate the intrinsic branch behavior of each important kernel in the H.264/AVC decoder. Fig. 7 shows the branch instruction distribution and mis-prediction rate for the whole decoder and individual kernels as well. We can see the the motion compensation (MC), entropy coding (EC), DBfilter are three

```

for (j = 0 ; j < BLOCK_SIZE ; j++) //BLOCK_SIZE=4
  for (i = 0 ; i < BLOCK_SIZE ; i++)
    for (result = 0, x = -2 ; x < 4 ; x++)
      // 6-tap FIR filter
      result +=imgY[max(0,min(y,y_pos+j))][max(0,min(x,x_pos+i+x))]*COEF[x+2];
    
```

Fig. 8. Fractional pixel interpolation in inter motion compensation

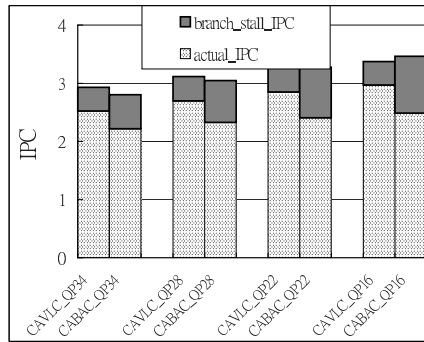


Fig. 9. CAVLC & CABAC branch stall comparison in VLC kernel

kernels that contain most branches. Below we discuss the branch behavior of these three kernels in details and show what kind of software optimizations can be done to improve branch prediction accuracy.

The motion compensation kernel contains nested loops as shown in Fig. 8. In the innermost loop, there are conditional branches associated with “min” and “max” operations. To reduce branch misprediction, we perform two types of optimizations. First, we unroll the innermost loop for 6-tap FIR filter. Second, we add two conditional instructions (conditional move when zero or not zero, CMOVZ and CMOVNZ) to eliminate conditional branches associated with “min” and “max” operations. These two optimizations can reduce 72% branch stall time in the motion compensation kernel.

The H.264/AVC standard provides two types of entropy coding, CAVLC and CABAC. CABAC provides more bitrate saving than CAVLC under the same visual quality. Fig. 9 compares the branch stall time between CAVLC and CABAC for different bit rates. We can see that the CABAC results in more branch stall time compared to the CAVLC. The CABAC method adopts arithmetic coding that processes one bit at a time. It depends on the bit value to determine the corresponding operation. This introduces a significant amount of unpredictable branches. The results also show that the sequence bit rate affects the amount of branch stall time in CABAC because the arithmetic coding has

to decode more symbols and the unpredictability property adds a lot of stalls when bit rate increased.

The deblocking filter of the H.264/AVC decoder smooths the undesired discontinuity between block edges and increases the subjective and objective visual quality of decoded pictures. The inputs to the filter process are pixels in adjacent blocks. The following condition is checked to determine whether filtering on the edge is necessary:

$|p_0 - q_0| < \alpha$ && $|p_1 - p_0| < \beta$ && $|q_1 - q_0| < \beta$, where p_0, p_1, q_0, q_1 represent pixel values.

The threshold values α and β are dependent on the average quantize parameters (QP) and two controlling thresholds (FilterOffsetA and FilterOffsetB) in the slice level. To reduce branch stall time caused by this condition checking, we use absolute instructions to eliminate branches associated with absolute value calculation. We find that the branch stall time of the DBfilter kernel is reduced by 40% using absolute instruction.

5 Conclusions

In this paper, we analyze the performance bottleneck of running the software implementations of the H.264/AVC decoder on modern superscalar processors. We find that the H.264/AVC presents a significant amount of instruction level parallelism. Unpredictable branch behavior appears to be the main performance bottleneck. Two new features, CABAC and deblock filtering, complicate control flows. In contrast to the commonly held belief, the H.264/AVC shows good cache performance despite of its streaming access pattern. This is because there exists significant data reuse at the macroblock level. In the future, we will study the effect of SIMD (Single Instruction Multiple Data) instructions.

Acknowledgements. This work is supported in part by the National Science Council under Grant NSC 93-2752-E-002-008-PAE and research funding from Microsoft. We would also like to thank the anonymous reviewers for their valuable comments.

References

1. "Draft ITU-T Recommendation H.264 and Draft ISO/IEC 14 496-10 AVC," in Joint Video Team of ISO/IEC JTC1/SC29/WG11 & ITU-T SG16/Q.6 Doc. JVT-G050, T. Wieg, Ed., Pattaya, Thailand, Mar. 2003.
2. T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 560-576, July 2003.
3. D. Marpe, H. Schwarz, and T. Wiegand, "Context-adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 620-636, July 2003.

4. H. Malvar, A. Hallapuro, M. Karczewicz, and L. Kerofsky, "Low-Complexity transform and quantization in H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 598-603, July 2003.
5. P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 614-619, July 2003.
6. Joint Model Version 7.3 (AVC/H.264 Reference software). Available via <http://bs.hhi.de/~suehring/tml/>
7. Horowitz, M. Joch, A. Kossentini, F. Hallapuro, A., "H.264/AVC baseline profile decoder complexity analysis" *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 704- 716, July 2003.
8. Matthew J. Holliman, Eric Q. Li, and Yen-Kuang Chen, "MPEG Decoding Workload Characterization"
9. D. Burger and T.M. Austin, "The SimpleScalar Tool Set, version 2.0," Technical Report 1342, Computer Science Dept., Univ. of Wisconsin-Madison, 1997.

An Efficient Traffic Smoothing Method for MPEG-4 Part-10 AVC/H.264 Bitstream over Wireless Network

Kwang-Pyo Choi, Inchoon Choi, Byeungwoo Jeon, and Keun-Young Lee

School of Information and Communication Engineering, Sungkyunkwan University
300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, Republic of Korea.
{sinawe, sonne, bjeon, kylee}@ece.skku.ac.kr

Abstract. In this paper, we propose an efficient traffic smoothing method for MPEG-4 Part-10 AVC/H.264 bitstream over wireless network. In general, the conventional traffic smoothing schemes try to reduce network congestion, however the main target of the proposed method is to reduce artifacts such as frame skip, occurred at the receiver. The proposed method consists of packet-based windowing and probing method. It can be efficiently applicable to the slice structure of H.264, and reduce computation complexity. In our experimental results, the proposed method shows promising results, which maximizes the degree of traffic smoothing and minimizes the frame skip count of receiver adaptively.

1 Introduction

There are many applications to rely on the transmission of live or pre-recorded video. Although video compression techniques such as MPEG-4 can effectively reduce the storage and the bandwidth requirement of channel, the transmission of compressed video typically causes significant burstness on a short time interval, due to natural variations within and between scenes. This variation of the bandwidth requirements complicates the transmission of video bitstream. Especially, in case of statistical multiplexing channel, network congestion often occurs and channel error is propagated. Therefore, the reduction of burstness and the peak bandwidth requirements is important problems in the transmission of video bitstream. One simple way to solve the problem is to use a traffic smoothing, which maintains the traffic of network at an average rate.

One example of a traffic smoothing technique is work-ahead smoothing [1], which uses a *priori* knowledge of pre-coded media. In this method, bitstream is sent according to a transmission schedule that generates a series of constant bit rate (CBR) segments, and a receiver stores it in a temporary buffer in advance of playback. However, although the performance of this method is optimal at a given the receiver's buffering time, it is not suitable to the heterogeneous network having different receiver's buffering time.

The mismatch between the receiver buffering time and the actual receiver buffering time results in *frame skip*, due to delay of bitstream. In other words,

if a packet does not arrive at the receiver having a small buffering time, the receiver cannot display the frame and the late frame should be skipped to prevent propagation of playing delay. In order to reduce this artifact without a *priori* knowledge of receiver's buffering time, we propose a traffic smoothing method using windowing and probing for H.264 bitstream.

This paper is organized as following, we briefly review a framework for media transmission over 3G network in Sect. 2. In Sect. 3, we explain H.264 for packet-based network. In Sect. 4, we illustrate the proposed method. In Sect. 5, we show the experimental results with respect to frame skip and traffic smoothing of the proposed method. And, we draw conclusion.

2 Media Transmission over 3G Network

The third generation (3G) of wireless communication system supports traditional voice communication and also evolves IP network to provide multimedia service. Especially, The General Packet Radio Service (GPRS) has been standardized by ETSI as a part of GSM Phase 2+ development. The phase 2+ specifications define the implementation of packet switching within GSM.

Packet switching means that GPRS radio resources are used only when users are actually sending or receiving data. Rather than dedicating a radio channel to a mobile data user for a fixed period of time, available radio resources can be shared between several users. The actual number of supported users depends on applications being used and the amount of data being transferred. Through multiplexing of several logical connections to one or more GSM physical channels, GPRS flexibly utilize channel capacity for an application with variable bit rates. This statistical multiplexing feature improves the efficiency of channel usage, however, it arises a problem that the transmission of video data is more difficult because a compressed video bitstream causes significant burstness on small time scales, which may result in network congestion or channel error. In order to an effective traffic smoothing method is required.

In case of 3G network, approximately 100 bytes of Maximum Transfer Unit (MTU) size is recommended because of its high bit error characteristic [2]. Comparing with the Ethernet network which uses 1500 bytes, it is very small. Although MTU size of 3G network is small, a video packet should be within MTU size in order to prevent fragmentation of it and be decoded independently. It suggests that we should use video coding scheme based on slice, which can encode a part of video frame in fixed byte size.

3 H.264 for Packet-Based Network

H.264 [3] is a new video coding technique developed jointly by the International Telecommunications Union (ITU) and the Motion Picture Experts Group (MPEG). As a ratified standard, it is known as MPEG-4 Part-10 AVC/H.264. H.264 provides a quite efficient mechanism for compressing and decompressing

motion video. This mechanism or algorithm requires significantly less bandwidth to transmit a video than what has been possible previously.

H.264 makes a distinction between a Video Coding Layer (VCL) and a Network Abstraction Layer (NAL). The output of encoding process is VCL data, which are mapped to NAL units prior to transmission or storage. The purpose of separately specifying the VCL and NAL is to distinguish between coding-specific features (at the VCL) and transport-specific features (at the NAL).

The method of transmitting NAL units is not specified in the standard but some distinction is made between transmission over packet-based transport mechanisms such as packet-based networks and transmission in a continuous data stream such as circuit-switched channels. In a packet-based network, each NAL unit may be carried in a separate packet and should be organized into the correct sequence prior to decoding. It is possible to use a transport mechanism such as the Real-time Transport Protocol (RTP) [4] to achieve sequencing of separate packets.

H.264 recommends that a slice should be packetized into a RTP packet within MTU size. It means that it would prevent a slice from being fragmented as already mentioned in Sect. 2. Therefore, the proposed traffic smoothing method considers MTU size constraint of 3G network and slice structure of H.264.

4 Proposed Method

4.1 Packet-Based Traffic Smoothing

In the conventional traffic smoothing method, a compressed video bitstream is arbitrarily separated into several packets after traffic scheduling because it concentrates on only optimal smoothing rather than the syntax of the bitstream [1]. Therefore, a receiver must receive all separated packets to reconstruct video correctly. It is not an efficient method because a compressed video bitstream cannot be reconstructed even in case that only one packet has been lost or late for display time of related frame.

In order to solve the problem, a suitable position to separate bitstream may be aligned at slice boundary. Of course, a slice with big size must be fragmented into several packets to smooth traffic efficiently. However, according to the recommendation of NAL in H.264 for RTP transmission and the MTU size of the 3G network, the RTP packet size of a slice for 3G network should be very small. Therefore, if we use traffic smoothing based on the small RTP packet rather than arbitrary byte size for segmentation, the problem mentioned above may not be occurred and it even simplifies transmission system, even though optimal traffic smoothing could not be achieved. Figure 1 illustrates the difference between the traffic smoothing methods, i.e. byte-oriented traffic smoothing and packet-based traffic smoothing.

4.2 Frame Skip Caused by Traffic Smoothing

Already mentioned, a traffic smoothing technique is effective to reduce network burstness in video transmission system. However, an excessive traffic smoothing

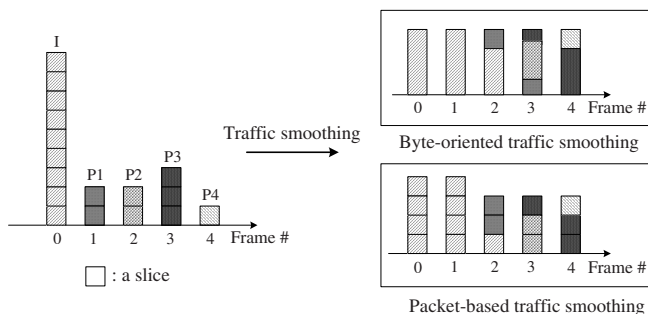


Fig. 1. Byte-oriented and packet-based traffic smoothing. I and P1~P4 mean Intra and Inter coded video bitstreams, respectively.

causes additive delay at a receiver. In other words, the receiver must wait for the delayed packet to reconstruct the video because traffic smoothing is the technique which adds intentional delay to compressed video bitstream to reduce network burstness.

Therefore, if delayed packets have not been arrived at the receiver before display time, the display might be paused until all remained packets are received. But then, even though the delayed packets are received and video frame is reconstructed correctly, the receiver cannot play the reconstructed frame at the moment because to play the reconstructed frame which misses its playing time occurs a permanent delay and the receiver buffer could overflow in worst case. After all, in such cases, the frame should be skipped at decoder, which occurs degradation of subjective quality. Therefore, the excessive traffic smoothing should be avoided in order to preserve the subjective quality of video.

As a summary, basically, the amount of traffic smoothing is not independent of frame skipping at the receiver and the overflow of the receiver buffer. The traffic smoothing technology should be properly designed so that it reduces network burstness and frame skipping at the receiver simultaneously.

4.3 Packet-Based Windowing and Probing

We introduce a new traffic smoothing method, called Packet-based Windowing and Probing (PWP), which is shown in Fig. 2.

This method uses packet-based windowing and probing mechanism to gather a receiver’s statistics via feedback channel such as Receiver Report (RR) of Real-time Transport Control Protocol (RTCP) [4] to decide traffic smoothing window size. The traffic smoothing window represents the amount of video frame count to apply traffic smoothing. For example, in Fig. 1, the window size is 5. Firstly, a sender decides an initial window size w_{init} , and builds segments to send by using packet-based traffic smoothing as already shown at Sect. 4.1, and transmits the segments. And then, the receiver reconstructs and displays the received segments. If a frame skip is occurred at the receiver, the amount of frame skip is counted and

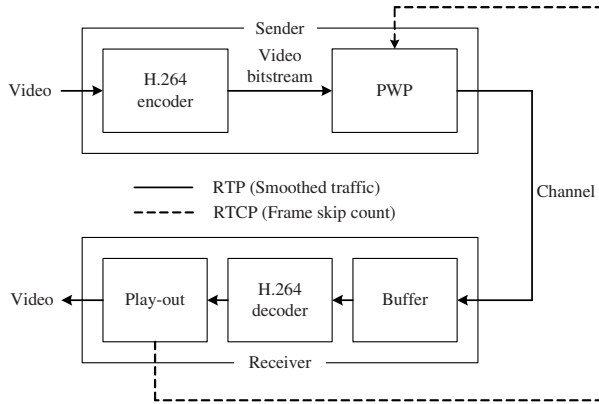


Fig. 2. Block diagram of the proposed method, Packet-based Windowing and Probing

sent to the sender periodically by using RR of RTCP containing the cumulated number of frame skip. If the sender receives RR of RTCP, the sender calculates the new window size \hat{w} by using the current window size w and the number of frame skip s according to the following equations;

$$\text{If } (s < \theta) \quad \hat{w} = \min(w + \alpha, w_{\max}), \quad \alpha = 1, 2, \dots, w_{\max}, \quad (1)$$

$$\text{else } \hat{w} = \max([w \cdot \beta], w_{\min}), \quad 0 < \beta \leq 1, \quad (2)$$

where, $[\cdot]$ implies round off operation. θ , w_{\max} , w_{\min} , α , and β are threshold for decision, maximum window size, minimum window size, increment factor, and decrement factor respectively. If the frame skip count s is smaller than the predefined threshold value θ , the window size is increased by adding increment factor α to increase traffic smoothing effect. Otherwise, the window size is decreased by multiplying by decrement factor β to prevent the frame skip possibility at the receiver. In (2), we used multiplication rather than subtraction to prevent the successive frame skip in next transmission until next window size is decided.

After new window size \hat{w} is calculated, the sender performs traffic smoothing according to the new window size. In traffic smoothing process, the sender must know how much bandwidth it occupies at once. The bandwidth can be calculated by simply averaging the size of data in a window size. However, as already mentioned in Sect. 4.1, it must be aligned in packet boundary. The following recursive formed a equation is method to obtain sending RTP packet count c_i of i -th segment according to the count of packet of j -th video frame, f_j in a window w ;

$$c_i = \left\lceil \frac{1}{w-i} \left(\sum_{j=0}^{w-1} f_j - \sum_{j=0}^{i-1} c_j \right) \right\rceil \quad i = 0, 1, 2, \dots, w-1, \quad (3)$$

where, $\lceil \cdot \rceil$ implies round up operation. Above windowing and probing sequences are continued until the end of transmission.

Table 1. The conditions for experiments

Video coding (H.264 baseline)	Video sequence	Foreman, CIF, 300 frames, 30Hz
	Intra interval	None (BS 1) and Every 15 frames (BS 2)
	Slice mode	80 bytes fixed
	Deblocking filter	Off
	Rate control	Disable
PWP	Parameters	$\alpha = 2, \beta = 0.5, \theta = 1, w_{init} = 2$
	RTCP period	When a new window size is calculated

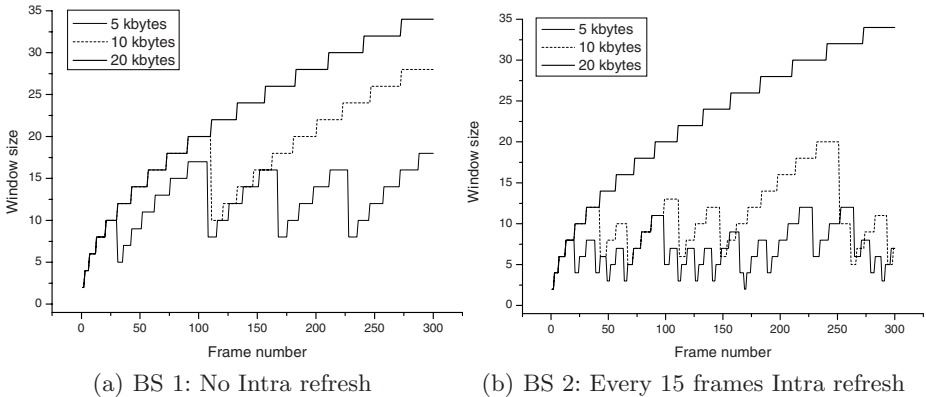


Fig. 3. Variation of window size with respect to receiver buffer size.

5 Experimental Results

The proposed method, PWP is simulated with respect to various receiver buffer size. We used JM7.5 to encode 300 frames Foreman video sequence, which is often used to evaluate coding efficiency of video codec. Two types of Intra refresh intervals are used. One (BS 1) is an encoded bitstream which does not contain periodic Intra refresh, the other (BS 2) is one which has Intra refresh frame in every 15 frames. The detailed conditions are shown in Table 1¹.

5.1 Window Size with Respect to Frame Skip

The variation of window size with respect to frame skip is shown in Fig. 3.

As shown in Fig. From the first video frame, window size is increased because frame skip count s does not exceed threshold value θ . At each dropping position,

¹ Usually, RTCP period is larger than that of Table 1. However, to evaluate the proposed method by using 300 frames Foreman sequence, we sent RTCP whenever a new window size would be calculated.

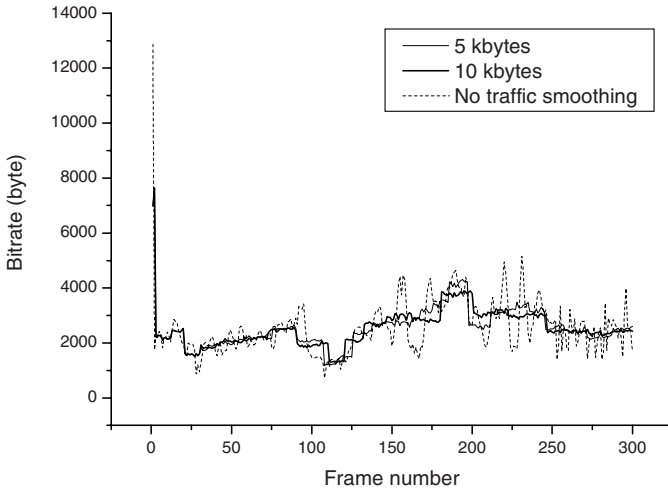


Fig. 4. Traffic smoothed bitrates of BS 1 (No Intra refresh).

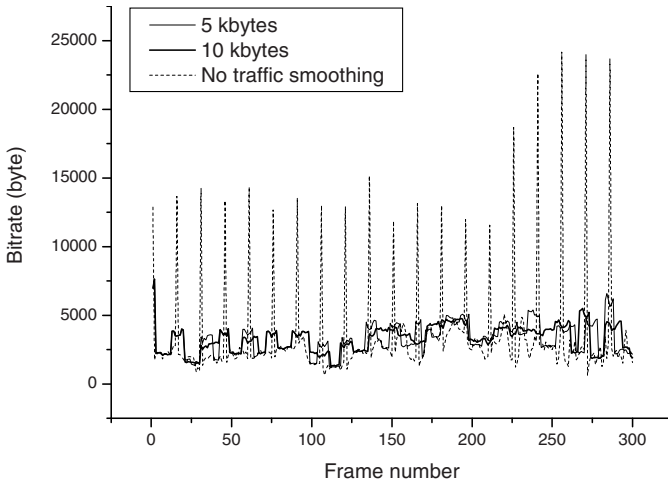


Fig. 5. Traffic smoothed bitrates of BS 2 (Every 15 frames Intra).

window size is rapidly decreased due to frame skip count exceeding the threshold value. And then, window size is increased again to bring up the traffic smoothing effect. In Fig. 3, the receiver having buffer size of 20 kbytes does not show any frame skip in 300 frames. But If the video sequence has more frames than 300, it is expected that a frame skip would be occurred as the window size is additively increased.

5.2 Traffic Smoothing with PWP

Fig. 4 and 5 show original bitstream bitrates and traffic smoothed bitrates according to the proposed method, PWP.

Generally, the bitrate of video bitstream has a peak point when Intra frame occurs, which affects burstness of network. Using PWP, Intra frames and Inter frames are regulated within a window size and each peak point is smoothed over all frames. Comparing with Fig. 3, if larger size of window is used, the smoothing effects will be also increased. Therefore, the proposed method could be an efficient method to achieve both maximizing traffic smoothing and minimizing the frame skip count of receiver adaptively.

6 Conclusion

In this paper, we proposed a new traffic smoothing method, called PWP, using packet-based windowing and feedback probing for H.264. In 3G network with H.264, PWP method significantly eliminates the artifact of conventional traffic smoothing method and simplifies transmission system even though optimal traffic smoothing could not be achieved. As shown in several experiments, the proposed method shows the result of maximizing traffic smoothing and minimizing frame skip count of receiver adaptively. Therefore, our method could be efficiently used to smooth traffic for receiver having low-power in H.264 video coding and wireless network environments such as 3G network. For the sake of intensive evaluation of our method, it needs to experiment with respect to more complicated conditions and parameters. It will be a guide to improve the proposed method.

References

1. J. D. Salehi, Z.-L. Zhang, J. F. Kurose, and D. Towsley: Supporting Stored Video: Reducing Rate Variability and End-to-End Resource Requirements through Optimal Smoothing, *Proceedings of ACM SIGMETRICS*, May (1996) 222–231
2. G. Roth, R. Sjoberg, G. Liebl, T. Stockhammer, V. Varsa, and M. Karczewicz: Common Test Conditions for RTP/IP over 3GPP/3GPP2, ITU-T SG16 document VCEG-N80, Dec. (2001)
3. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC), JVT-G050r1, May (2003)
4. H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson: RTP: A Transport Protocol for Real-Time Applications, RFC 1889, IETF, Jan. (1996)

An Error Resilience Scheme for Packet Loss Recover of H.264 Video

Ziqing Mao^{1,*}, Rong Yan², Ling Shao², and Dong Xie²

¹ Computer science Dept, Tsinghua University, PRC,
mzq@mails.tsinghua.edu.cn

² IBM China Research Lab, 2/F, Haohai Building, No.7, 5th Street, ShangDi,
Beijing, 100085, PRC,
yanrong@cn.ibm.com

Abstract. In this paper, we propose both an error resilience scheme and the corresponding redundant slice selection and error concealment techniques for the H.264 bit stream transmission over packet-switching networks. To recover the degradation from packet loss, we use redundant picture coding in the encoder side as well as motion compensated error concealment for the decoder. Thereon, a redundant slice selection algorithm is invented to make redundant picture coding efficient. Meanwhile, a well-designed motion vector prediction method is proposed to get precise motion vectors for lost macroblocks. The experimental results for different sequences with diverse packet loss rate show that the proposed error resilience scheme works well for packet loss recovering. The improvement of objective quality is distinguished, e.g. up to 5.8dB for whole sequence and up to 7dB for chosen segment, as well as that of subject experience.

1 Introduction

Many efforts have been put on the development of effective techniques for packet loss resistance for video transmission over networks. In general, error resilient approaches include three categories [1], namely the error resilient encoding approach [1], the error concealment approach [2][3], and the encoder-decoder interactive error control approach.

Error concealment at the decoder is to estimate the missing samples from the correctly received neighboring ones, which are spatial or temporal. In some cases, e.g. the information of spatial and temporal neighbors are not available; or, the video content of the missing sample and that of its spatial and temporal neighbors are very different, the concealed results are usually poor. So some error resilience tools are used in encoder to facilitate the error concealment on the decoder side.

The H.264 provides a toolkit of error resilience [4] including redundant coding picture, flexible macroblock ordering (FMO), random intra macroblock (MB)

* This work was done when author was with IBM China Research Lab.

refresh, etc. The error resilience tools will bring extra redundancy to video signals. On the encoder side, several options can be turned on so that a trade-off between compression rate and error resilience is made targeting different type of problems found in heterogeneous environments [5].

Redundant coded picture is a coded representation of a picture or a part of a picture. When some of the samples in the decoded primary picture cannot be correctly decoded due to losses in transmission of the sequence and the coded redundant slice can be correctly decoded, the decoder should replace the samples of decoded primary picture with the corresponding samples of the decoded redundant slice [5]. Therefore, the slices which are coded in the redundant coded picture are much more reliable. Herein lie challenges to choose the samples should be reduplicated and to decide the quantity of redundancy. So in this paper, we first invent an algorithm for redundant slice selection. It makes redundant picture coding practically and efficiently.

After that, a MV prediction method named recursive average prediction (RAP) is proposed to estimate motion vector (MV) accurately for the lost MB. Besides, by concealing lost MBs in a snake-in order, it takes advantage of those redundant slices. As a result, the quality of concealed video is greatly improved.

In section 2, the redundant slice selection algorithm is described in details; also a set of criteria is discussed. Recursive average prediction is introduced in section 3. Till now, an whole error resilience scheme, with redundant picture coding in the encoder and motion compensated error concealment in the decoder, are proposed for H.264 video transmission being subject to packet loss. Simulation results, regarding the PSNR improvement brought respectively by redundancy and concealment, are presented in Section 4, which lead to the conclusions of this paper in Section 5.

2 The Algorithm for Redundant Slice Selection

Although the redundantly coded picture objectively reduces the packet loss rate of the replicated samples, it brings overhead. Therefore, the point is to develop an algorithm for redundant slice selection to trade off compression efficiency and error robustness.

2.1 Which Samples Need Redundancy

Intuitively there are two kinds of samples needing more protection. The first is the sample with large residual. Such sample is of less redundancy regarding their temporal/spatial neighbors. It usually contains new content never appeared in the foregoing sequence. Since almost all concealment algorithms use information from neighboring samples to recover the lost sample, the distortion of concealed sample is considerable. The other is the sample referred by large number of samples in the subsequent pictures. We call this as "expanding", where the distortion from the poor concealment would largely propagate in the following pictures.

Generally, the following four situations lead to large residual and expanding:

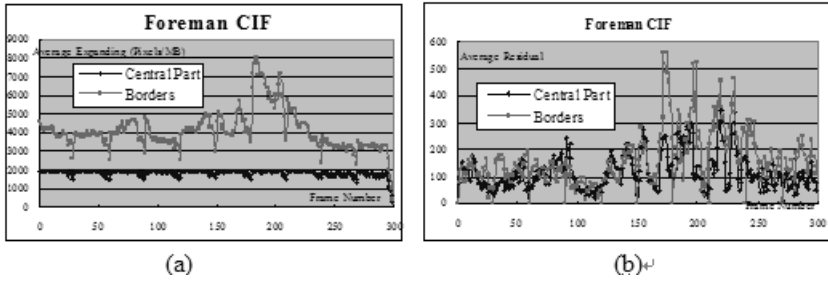


Fig. 1. (a)The average expanding&(b)the average residual of borders and central part

1. Sudden: The scene is relatively static, and a new content appears somewhere or an object suddenly expands.
2. Move-into: A moving object moves into the visual field from outside gradually. The moving-into object always brings new content; also the expanding of samples covers the moving object.
3. Pan: The camera is panning, the new content appears and expands along the direction of panning.

4. Switch: The sequence switches and the whole picture contain new content.

The "sudden" is stochastic and the new content may appear anywhere. It is difficult to estimate effectively. The "switch" will not be processed separately in this paper. The "Move-into" and "Pan" happen quite often, where the new contents always appear and expand along the edges of the pictures. We use following statistic data to prove above conclusions.

The Fig 1.(a) is the statistic result of average expanding of Foreman CIF. Each "border" consists of the macroblocks on one edge of the picture; there are four borders in one frame, denoted as G_1 , G_2 , G_3 and G_4 . The other MBs are referred as the "central part" in Fig. 2. The "expanding" of a macroblock is the number of the pixels referring to it in the subsequent pictures, e.g. 5 pictures here. The "average expanding" of "borders" is calculated by averaging the expanding of all the macroblocks belong to the "borders". "Average expanding" of "central part" is calculated in the same way by considering MBs of "central part". From the statistic data, we can see that the expanding of the borders are remarkable larger than that of the central part. The curve decreases from frame 295 because there are totally 300 frames in the test sequence.

The average residual of the borders and the central part for the same sequence are shown in Fig. 1.(b). Average residual is the average of prediction errors of all macroblocks in the corresponding part that is either "borders" or "central part". Seen from the figure, most average residual of borders is larger than that of central part. Although the gap between them is not so large on average, the macroblocks with considerable residual, e.g. more than 300 in Fig. 1. (b), mainly belong to the borders.

According to above analysis, it is the borders that to be coded redundantly for their considerable residual and large expanding.

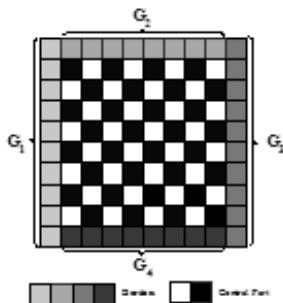


Fig. 2. The concept of the border and the central part

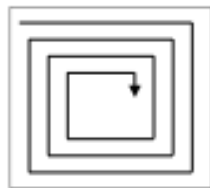


Fig. 3. The "snake-in" order for the concealment

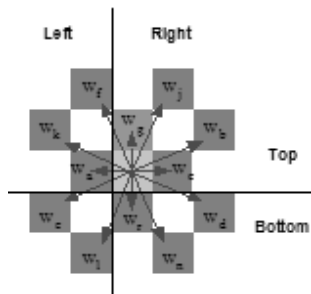


Fig. 4. The prediction window centered in the missing macroblock

2.2 The Redundant Slice Selection Algorithm

We in this paper encode one H.264 picture into several slices.

A primary redundant slice selection algorithm according to above analysis is to code the borders as a redundant slice. However, the overhead of this method is too large. The statistic result of Foreman sequence shows that it is about 40% for QCIF and 30% for CIF. It is necessary to cut down the overhead and improve the efficiency of the redundant slices.

Since the new content and the expanding always occur along certain directions in "Pan" or "Move-into" case, a more efficient approach is to prioritize the borders by their residual and expanding effect then differentiate a specific one. So we further consider each border separately. That is if $r_i > r$ or $e_i > e$, the will be coded as a redundant slice, where $1 \leq i \leq 4$, r_i is the average residual of border G_i , e_i is the average expanding of group G_i , r and e are the pre-defined thresholds for the average residual and the average expanding respectively. The overhead can be balanced by adjusting the pre-decided threshold, r and e , for different sequences.

Seen from the experimental results in section 4, we can find that the proposed redundant slice selection algorithm enables the redundancy to be added for those samples really with large residual and expanding, so that the most benefits are brought while little overhead are introduced. It not only does well in the "Pan" and "Move-into" cases, but also to be beneficial for "Sudden" and "Switch" cases obviously. The analysis of experimental results is presented in section 4 in details.

3 Error Concealment Algorithm at the Decoder Side

To facilitate error concealment at the decoder side, besides redundant picture coding for the borders, we use the scatter slice partition to code the central part [6]. As shown in Fig. 2., each macroblock and its four neighbors are filled in the

two different slices denoted by the white and the black. The followings describe the error concealment algorithm concretely.

3.1 Motion-Compensated Concealment in “Snake-in” Order

The proposed concealment algorithm is MB-based. Not only the correctly received but also the concealed macroblocks are treated as reliable neighbors in the concealment process. As shown in Fig. 3., the processing starts with the up-left macroblock and snake-in MB-by-MB. For the border is redundantly encoded and concealed first, more information is available for the central part.

3.2 Recursive Average Prediction

For each lost macroblock, motion compensated temporal concealment is performed. It is to “guess” the MV of the missing macroblock by some prediction schemes from available motion information of spatial or temporal neighbors. This “guessed” MV is then used for motion compensation using the reference frame.

Actually, there are several operations for estimating lost MV [1]. Setting lost MV as zero is the most popular one for it is simple, while using the MV info of spatial or temporal neighbors always reduces artifacts in the presence of large motion. To estimate lost MV, one can either use the MV of adjacent MB/block [2][7] or reconstruct the MV [7], e.g. weighting the MVs of spatial neighbors with a fixed weighing matrix. Such approach essentially assumes that the motion among those spatial MBs is in one trend and, which is more exactly, is approximatively consecutive. But the actual conditions are not all this case. It is possible that the motion of the top MBs is opposite to the motion of the bottoms, thus the average of motion in the top and bottom is unreasonable.

So we propose a new prediction algorithm called recursive average prediction (RAP) by dividing the spatial neighbors of the lost MB into different groups recursively, and only MVs from groups of approximatively consecutive motions are used as the candidates for prediction of recovered MV.

Consider a prediction window, comprising $m = n * n$ MBs, as shown in Fig. 5. The weighting matrix $\{w_i\}$, corresponds to the prediction window assigns a weight w_i to each available spatial neighbor, where $0 \leq i \leq m/2$. Firstly we divide the prediction window into two sub-windows along vertical direction (the top and the bottom as shown in Fig. 5.), and then calculate the averaged MV for each, MV_{top} and MV_{bottom} by the weighting matrix.

If the gap between the averaged motion vectors of two parts is smaller than a pre-defined threshold, it shows that the motion along the vertical direction is approximatively consecutive and the average of the two motion vectors is considered as a candidate. If the gap is larger than the threshold, the motion along the vertical direction is not consecutive. Then we use the same approach recursively in the top and bottom sub-windows respectively. The recursion terminates until current sub-window contains only one available macroblock and the motion vector of the only macroblock is considered as a candidate.

After above steps, secondly the same process is applied along the horizontal direction. The recursive prediction scheme actually can be summarized as:

1. Divide prediction window/sub-window into two sub-windows along the vertical direction;
2. Calculate MV_{top} and MV_{bottom} of these two sub-windows by the weighting matrix $\{w_i\}$;
3. If $|MV_{top} - MV_{bottom}| < \gamma$, then $MV_{recover} = ave(MV_{top}, MV_{bottom})$;
4. Otherwise, recursively execute the steps 1, 2 and 3 for the sub-windows, until only one MB in the current window;
5. Divide prediction window/sub-window into two sub-windows along the horizontal direction and perform above steps. The "top/bottom" is replaced by corresponding "left/right"; Fig. 5. shows a possible splitting of prediction window/sub-window.

3.3 Simplification of Recursive Average Prediction

This approach generates a set of candidates for the recovered MV. Each candidate essentially maps to a sub-window with approximatively consecutive motion. However, not every sub-window contributes evenly to the MV prediction. Those sub-windows with small size (e.g. contain only one macroblock) or apart far away from the lost macroblock contribute less to the final result. So we derive a simplified one from original recursive average prediction, named as SRAP, which calculates in advance all possible sub-windows benefit for the prediction instead of executing recursive process. Thus the recursive average prediction is made more efficient.

Fig. 5. illustrates an example of predefined multiple weighing matrixes. Each column of two weighting matrixes has different weighting center, e.g. middle, up, down, left and right, to approximate the recursive average prediction processing. It avoids the unwarranted assumption of applying a fixed weighted matrix and provides precise MV prediction. The experimental results show high performance of the approach that will be discussed in section 4.



Fig. 5. The multiple weighting matrixes

The decision that which candidate should be used as prediction MV for the missing macroblock is calculated by the boundary matching algorithm as introduced in [2][3].

4 Experimental Results

The simulations were carried out using H.264-JM7.3 reference code [8]. Three test sequences, Foreman, Coastguard and Flower are used to evaluate the performance of the proposed scheme. These sequences are chosen for their different motions and textures. The formats consist of CIF and QCIF. The group of picture (GOP) structure is used in encoding. Each GOP starts with one I frame followed by 29 P frames. The sequences are transmitted at packet loss rate (PLR) of 1%, 5%, and 10%. The PSNR of each decoded video sequence was calculated for 30 runs.

The proposed algorithms were compared to a simple and popular recovery method, temporal extrapolation, which conceals the errors by copying corresponding macroblocks from the previous frame [7]. In the following tables and figures, "TE" denotes the result of the temporal extrapolation method; "SRAP" denotes the results of error concealment where simplified recursive average prediction is adopted to calculate candidate MVs; "SRAP-Redundancy" denotes the result with both error concealment with SRAP and the redundant picture coding performing proposed redundant slice selection algorithm.

Table 1. presents the PSNR of different sequences for each case. For all the sequences, the proposed error recover scheme improves the quality of video averagely over 1dB and up to 5.8dB. For Foreman CIF, the improvements are 1dB-2.6dB for PLR 1%-10%. Therein, the benefit from proposed error concealment is about 0.7dB-1.9dB, the benefit from selected redundant slice is 0.3-0.7dB. Although the average improvement from selected redundant slice seems not so large as that from concealment, it is much better for those frames with more "Pan" and "Move-in", which will be discussed in the following paragraph. The results

Table 1. The comparison of PSNR values for different methods

Sequence	Foreman CIF			Foreman QCIF			Coastguard CIF			Flower CIF		
PLR	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
TE(dB)	25.96	29.91	33.88	26.70	29.46	33.51	25.34	28.78	32.56	20.78	23.77	30.60
SRAP(dB)	27.83	31.20	34.58	27.97	30.60	34.09	27.49	29.45	33.46	25.05	28.18	32.80
SRAP-Redundancy(dB)	28.52	31.57	34.86	29.16	31.40	34.43	27.83	29.72	33.54	26.58	29.68	33.44

Table 2. The bitrate and overhead of each sequence

Sequence	Foreman CIF	Foreman QCIF	Coastguard CIF	Flower CIF
Original Bitrate	561.0kbps	183.0kbps	1199.8kbps	1715.2kbps
Bitrate with redundancy	597.9kbps	200.5kbps	1222.9kbps	1898.1kbps
Overhead	6.2%	8.7%	1.9%	9.6%

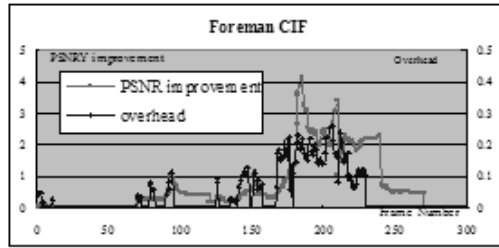


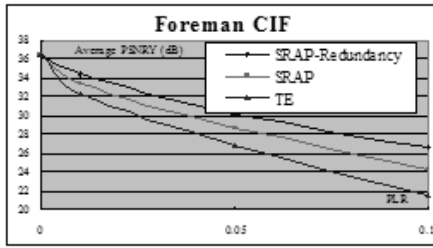
Fig. 6. The PSNRY improvement & overhead brought by redundancy for each frame

of Foreman QCIF show improvements reaching 1dB-2.5dB, for Coastguard CIF 1dB-2.5dB, and for Flower CIF 2.8dB-5.8dB. Furthermore, the improvements increase with the rise of PLR. It is noticed that since almost all the fames of Flower are "Pan", the redundant slices are more useful for it, which results in up to 1.5dB gain. The overhead brought by redundant slice at current bit rate is listed in Table 2. The overhead is calculated by $(b_r - b_o)/b_o$, where b_r is the bitrate with redundancy and b_o is the original bit-rate. It ranges from 1.9%-9.6% for different sequences, which is acceptable.

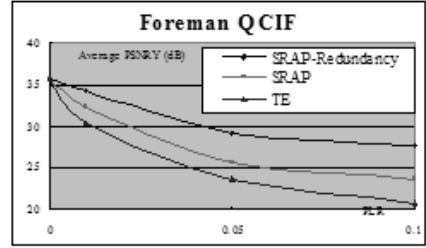
Actually, the redundancy is not evenly distributed across different frames within one sequence. By proposed redundant slice selection algorithm, comparing the "overhead" curve in Fig. 6. and "borders" curve in Fig. 1., the redundancy focuses more on those frames with relatively large residual and expanding. This also results in remarkable improvement of PSNR for them, which are named as "chosen segment" in the followings. As shown in Fig. 6., for Foreman CIF, when PLR is 10%, the average overhead of frame 180-240 is obviously larger than that of others; the PSNR improvement is relatively large.

The Fig. 7. illustrates the quality improvements for chosen segment of each sequence. For Foreman CIF, the improvements are 2.2dB-5.2dB for PLR 1%-10%, where the benefit from proposed error concealment is about 1.2dB-2.8dB, and from selected redundant slice is 1dB-2.4dB. Compare to the result of the whole sequence, redundant slice selection contributes more to the decoding quality for chosen segment and also enables more improvement from error concealment. Actually, the results of chosen segment of all test sequences are obviously better than that of the whole as shown in Fig. 8. The PSNR improvement for the "chosen segment" is up to 7dB, and the benefit from redundant slice selection is up to 4dB. It shows that the proposed redundancy coding is an efficiency method to solve the problems when pure error concealment fails to get acceptable result in the "Pan" and "Move-into" conditions. Although the discussion focuses on "Pan" and "Move-into" cases, it is to bring benefits in "Sudden" and "Switch" cases obviously.

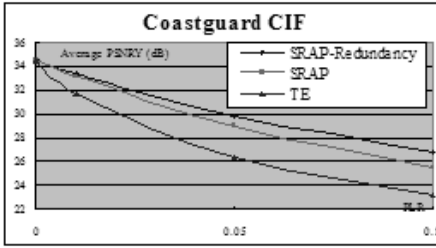
Finally, it is noted that not only the objective quality are improved significantly, the subject experience also improved by using the proposed error resilience scheme, which can be seen in Fig. 8.



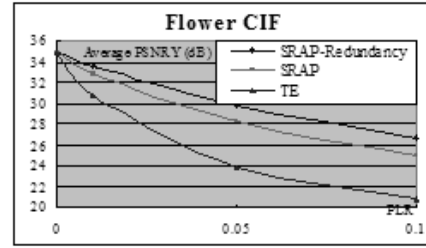
(a) Chosen segment: frame #180 - #240.



(b) Chosen segment: frame #270 - #330



(c) Chosen segment: frame #0 - #90.



(d) Chosen segment: frame #0 - #194.

Fig. 7. The PSNR improvement & overhead brought by redundancy for each frame

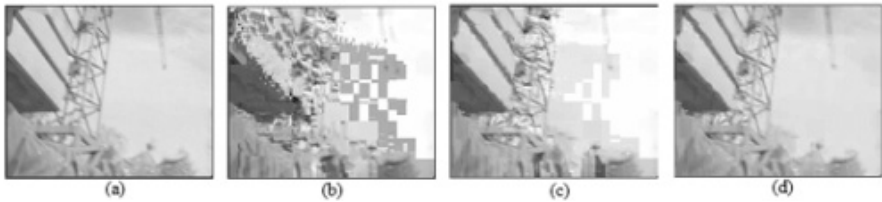


Fig. 8. The subjective experience of frame 191 in Foreman CIF: (a). Error-free; (b) TE; (c) SRAP; (d) SRAP-Redundancy

5 Conclusions

We propose in this paper an error resilience scheme as well as the corresponding techniques for packet loss recover of the H.264 bit stream.

In proposed scheme, motion compensated error concealment is deployed at the decoder side; at the same time the borders of one frame are selectively repeated for redundant picture coding and the central part is encoded in scatter slice partition mode.

To perform motion compensated error concealment effectively, the MBs are concealed in a well-designed snake-in order. Besides, a new recursive average prediction method is discussed and its simplified one is deployed. By this method,

the candidates for recovered MV are derived from MVs corresponding to those prediction windows/sub-windows with approximately consecutive motion, so more accurate MV is estimated for the lost MB.

However, pure error concealment fails to get enough good result when redundancy, in original pictures, between the missing sample and its neighbors is small, and especially it fails to get acceptable result when the error propagation brought by reference is relatively large. To solve the problems, we propose a novel algorithm of redundant slice selection for redundant picture coding. By the algorithm the samples with considerable residual and is referred by relatively large samples in the subsequence are redundantly encoded to overcome the "Pan" and "Move-into" cases, which the pure error concealment fails to deal with. And only little overhead is introduced into each sequence.

The experimental results show high performance of the proposed scheme. It improves significantly both subject and object quality of H.264 video when packet loss occurs. For each sequence, the improvements increase with the rise of PLR. The PSNR gain is up to 5.8dB for the whole sequence and 7dB for the chosen segment.

References

1. Y. Wang, Q. F. Zhu: "Error control and concealment for video communication: a review", Proceedings of the IEEE, vol. 86, pp. 974-997, 1998.
2. Y. Wang, M. M. Hannuksela, V. Varsa, A. Hourunranta, M. Gabbouj: "The error concealment feature in the H.26L test model", IEEE Int. Conference on Image Processing, pp.729-732, Rochester, NY, USA, 2001.
3. L. Atzori, F. G. B. De Natale, C. Perra: "A Spatio-temporal concealment technique using boundary matching algorithm and mesh-based warping (BMA-MBW)", IEEE Trans. Multimedia, vol. 3, Issue 3, pp.326-338, Sept. 2001.
4. T. Wiegand, G. J. Sullivan, G. Bjontegaard, A. Luthra: "Overview of the H.264/AVC video coding standard", IEEE Trans. CSVT, vol. 13, no. 6, pp.560-576, Jul. 2003.
5. "Text of ISO/IEC FDIS 14496-10: Information Technology - Coding of audio-visual objects - Part 10: Advanced Video Coding", ISO/IEC JTC 1/SC 29/WG 11 N5555, Mar. 31, 2003.
6. JM73.zip, <http://bs.hhi.de/suehring/tml/download/>, Aug. 2003.
7. J.Suh, Y. Ho, "Recovery of motion for error concealment", Proceedings of the IEEE TENCON99, vol. 1, pp. 750-753, Sept. 1999.

Key Techniques of Bit Rate Reduction for H.264 Streams

Peng Zhang, Qing-Ming Huang, and Wen Gao

Institute of Computing Technology, Chinese Academy of Science, Beijing, 100080,
China

{peng.zhang, qmhuang, wgao}@jdl.ac.cn, <http://www.jdl.ac.cn>

Abstract. In previous techniques of bit rate reduction transcoding, reusing the mode of the input MB is widely adopted. However, directly re-using the mode of input MB will cause additional losses for H.264 bit-rate scaling. Variable-size block-matching based motion estimation makes H.264 much more efficient than other coding standards while increasing its complexity. As the target bit rate changes, the best mode of the MB may also change. In this paper we concerned two key techniques for H.264 bit rate reduction transcoding: Mode decision and rate control. A fast MB mode decision algorithm and an improved rate control algorithm depending on statistics of input streams is proposed.

1 Introduction

Bit-rate scaling is a key solution to media adaptation in hybrid networks. Cascaded decoder-encoder is a straightforward approach which can achieve the best quality at the cost of high complexity. In the literature, some papers proposed different techniques to get a trade-off between acceptable quality and complexity over the past few years in which the focus has been centered on two specific aspects, complexity and drift compensation [1-4]. Generally, two architectures, close-loop and open-loop, are widely in use. Because open-loop systems are relatively simple but subject to drift-error, close-loop systems are more common in practical. [1] proposed an close-loop transcoder which requires one DCT and one IDCT, while cascade decoder-encoder requires one DCT and two IDCT. This architecture achieved nearly the same quality as cascaded decoder-encoder with some arithmetic inaccuracy introduced. All of these transcoding schemes are based on the assumption of fixed MB size for motion compensation. This assumption is effective to many coding standards, such as MPEG-2, MPEG-4 and H.263.

Variable-size block-matching (VSBM) based motion compensation is a key feature of H.264/AVS [5] which ensures its high coding efficiency. [6] proposed a Lagrangian mode decision technique by looping all available mode to select one with minimized rate-distortion cost. It is a computationally heavy procedure. However, directly reusing the mode of the input macroblock (MB) will cause severe quality losses because the best mode varies when its QP changes. Table-1

Table 1. Mode variations for different QP values, FOREMAN, CIF, 60 frames (ROW: QP = 32; COL: QP = 28)

	SKIP	P-16X16	P-16X8	P-8X16	P-8X8	I-4X4	I-16X16
SKIP	4222	658	80	80	7	10	67
INTER-16X16	1933	3155	452	470	97	56	96
INTER-16X8	347	783	696	200	148	27	33
INTER-8X16	525	1006	234	921	151	36	34
INTER-8X8	148	624	535	597	1291	64	10
INTRA-4X4	26	107	79	80	51	2071	394
INTRA-16X16	158	53	14	10	3	101	820

demonstrates the variation of mode at different QP, where the column is the mode distribution at the QP 32 and the row is at QP 28. From Table-1, we can see that the best mode may change as the QP value changes. In general, in the case of bit rate reduction transcoding for H.264 streams, the first step is to decide a proper QP for the target rate, then decision of MB mode for MC must be made. In this paper, two key techniques, rate control and mode decision, are concerned.

This paper is organized as follows: in section II, a mode classification approach is introduced and simplified rate-distortion optimization based on this mode classification is proposed. In section III, we propose a new rate control algorithm which fully utilizes the information from the input stream. Section IV are experimental result of our transcoding scheme and two comparison transcoders.

2 Mode Classification for Bit-Rate Scaling Transcoding

As we know, MPEG-2 [7] also supports variable block size motion compensation. Macroblocks of P frame can be coded in INTRA, INTER-16x16 or SKIP mode, and INTER-16x8 mode for interlaced videos. The small mode set makes it unnecessary to re-select modes in transcoding. H.263 [8] specifies INTER-8x8 coding mode. H.263+ adds some improvements in compression efficiency for the INTRA macroblock mode. Motion compensation in MPEG-4 [9] is based on 16x16 blocks and support INTER-8x8 mode. Also MPEG-4 includes alternate scan patterns for horizontally and vertically predicted INTRA blocks. However, the mode sets of these coding standards are rather small, and the reuse of the previous mode in transcoding has little effect on the quality.

H.264 supports variable prediction mode for motion estimation. For I frames, macroblocks can be coded in INTRA-16x16 or INTRA-4x4 modes, and furthermore, there are 9 prediction directions for INTRA-4x4 mode. For P frames, mode of target macroblock must be selected from SKIP, INTER-16x16, INTER-16x8, INTER-8x16 and INTER-8x8 mode respectively besides INTRA-16x16 and INTRA-4x4 modes. In case the INTER-8x8 macroblock mode is chosen, each 8x8 block can be further partitioned into blocks of 8x8, 8x4, 4x8 or 4x4

luminance samples. B frames are similar to P frames. Also, H.264 support multi-frame motion-compensated prediction. When bit-rate scaling, the best mode and best reference picture of target macroblock under the new rate may differ with those under its original rate. To maintain the quality, we classified the modes into different levels.

We know that the Lagrangian coder control will assign more INTER-8x8 or INTRA-4x4 mode to active regions and SKIP or INTER-16x16 to static backgrounds. That means active MB' mode level is high while static MB's mode level is low. So, we ordered the modes from low level to high level as SKIP, INTER-16x16, INTER-16x8, INTER-8x8 and INTRA-4x4. In the case of QP increases, the possible mode set follows the rule that the output mode order will not higher than the input mode order. For example, if the input mode is INTER-16x16, when the QP increases, the new best mode will be selected from modes whose level is lower than INTER-16x16. So, we classified all these modes into five classes for bit-rate scaling transcoding: 1.{SKIP, INTER-16x16};2.{INTER-16x8, INTER-8x16};3.{INTER-8x8};4.{INTRA-4x4};5.{INTRA-16x16}.

Based on this classification, we propose our mode decision algorithm. This mode decision algorithm decide the possible mode options for RDO which effectively restricts the Lagrangian method loops in a small range. Detailed algorithm is like the following pseudo-codes:

```

Switch( input macroblock mode)
{
case SKIP:
    Target mode set = {SKIP, INTER-16x16};
case INTER-16x16:
    Target mode set = {SKIP, INTER-16x16};
case INTER-16x8:
    Target mode set = {SKIP, INTER-16x16, INTER-16x8};
case INTER-8x16:
    Target mode set = {SKIP, INTER-16x16, INTER-8x16};
case INTER-8x8:
    Target mode set = {SKIP, INTER-16x16, INTER-16x8,
                      INTER-8x16, INTER-8x8};
case INTRA-4x4:
    Target mode set = {INTRA-4x4};
case INTRA-16x16:
    Target mode set = {INTRA-4x4, INTRA-16x16, SKIP};
}

```

After mode decision, different motion compensation scheme is used for different modes. If one MB's mode is decided as INTER-16x16, MV can be reused and traditional transcoding techniques can be directly adopted. Otherwise, if target modes includes INTER-16x8, INTER-8x16 or INTER-8x8, motion vectors refinement is needed. Because in most sequences, SKIP and INTER-16x16 modes possess a large percent, the above algorithm can efficiently reduce the

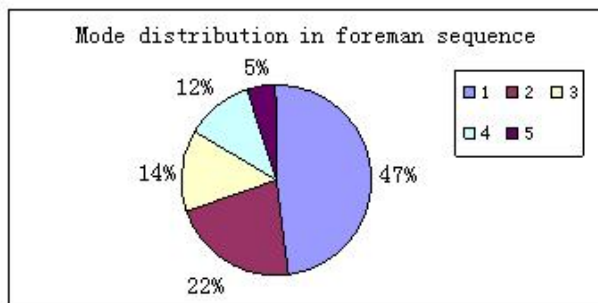


Fig. 1. 1: INTER-16x16 and SKIP mode, 2: INTER-16x8 and INTER-8x16 mode; 3: INTER-8x8 mode, 4: INTRA-4x4, 5: INTRA-16x16 mode

computation complex while maintaining the acceptable quality. Furthermore, to keep the quality of areas with high activity, in case the input macroblock mode is INTRA-4x4, the target macroblock mode is INTRA-4x4.

Figure 1 illustrates the computational complexity participation in the case of Table-1 as QP equals to 28. From figure 1, we can see that about 65% macroblocks are INTRA or INTER-16x16 modes which doesn't require RDO and motion estimation and only 22% macroblocks' mode can be selected from INTER-16x8, INTER-16x16 or INTER-8x16, INTER-16x16, and only 14% macroblocks require looping all INTER modes. So, the computational complexity is greatly reduced by this method. Also, this classification can be helpful in INTRA-Refresh technique for error-resilient transcoding to convert macroblocks with high activity into INTRA modes.

3 Rate Control Algorithm for H.264 Transcoding

Video encoding rate control is important to achieve consistent video quality and has been the interest of research in recent years [10-13]. As we know, the key role of R-Q model is to determine the QP before encoding one frame by available channel bandwidth, output buffer fullness, and picture and motion complexity. There are two problems in rate control: the first one is target frame bits allocation, and the second is the estimation of quantization parameters. [12] proposed and improved method named "MAD ratio" for target frame bits allocation. And in [14], the author proposed a two-pass QP prediction algorithm based on TM5. In video transcoding, the statistics such as number of bits and QP can be easily derived from the stream. We can compute the complexity of the picture and target bit allocation based on these statistics. [15] proposed a bit allocation algorithm based on the fact that the ratio of the complexity between input and output pictures keeps constant. However, this algorithm needs the information of the whole GOP, and is not fit for the sequences in which the content changes severely. In this paper, the target frame bit number depends on the actual coded bits number of the input frame and ratio of target rate

and input rate, and then QP can be predicted based on the first-order R-Q model. This rate control method can be adopted in frame level, object level and macroblock level.

In our rate control algorithm, the fluid traffic model is applied the same as that in [6]. Let $R_i(j)$, $B_i(j)$, $V_i(j)$ and $b_i(j)$ denote the instant available bit rate, total bits for the rest pictures, the occupancy of the virtual buffer and the actual generated bits in j^{th} picture respectively, where i and j means j^{th} picture in i^{th} GOP to be coded. And N_i denotes the total number of pictures in i^{th} GOP, f is the frame rate.

$$\begin{cases} \frac{R_i(j)}{f} \times N_i - V_i(j) & j = 1 \\ B_i(j-1) + \frac{R_i(j)-R_i(j)}{f} \times (N_i - j + 1) - b_i(j-1) & j = 2, 3, \dots, N \end{cases} \quad (1)$$

$$\begin{aligned} V_i(1) &= \begin{cases} 0 & j = 1 \\ V_{i-1}(N_{i-1}) & other \end{cases} \\ V_i(j) &= V_i(j-1) + b_i(j-1) - \frac{R_i(j-1)}{f} \quad j = 2, 3, \dots, N \end{aligned} \quad (2)$$

The determination of target bits for current P frame is composed of four steps as the following:

Step 1: Compute the target bits, this step can be divided into 2 or 3 sub-steps depends on whether frame level or basic unit level rate control is applied.

Sub-step 1: Determine target buffer level for current P picture

$$\begin{aligned} S_i(2) &= V_i(2) \\ S_i(j+1) &= S_i(j) - \frac{S_i(2)}{N_{p(i)}-1} \quad j = 2, 3, \dots, N_i \end{aligned} \quad (3)$$

Sub-step 2 Compute the target for current P picture The target bits allocated for the j^{th} P picture is determined based on the target buffer level, the frame rate, the available channel bandwidth, and the actual buffer occupancy as follows:

$$T_{buf} = \frac{R_i}{f} + \gamma \times (S_i(j) - V_i(j)) \quad (4)$$

Meanwhile, the remaining bits are also computed as:

$$T_r = \frac{B_{org}(j)}{N_j \times (R_{org}/f) - \sum_{k=0}^{j-1} b_{org}(k)} \times B_i(j) \quad (5)$$

Where $b_{org}(j)$ and $R_{org}(j)$ are actual input bits of the j^{th} picture and the channel bandwidth of the input stream. The final target bit T is a weighted combination of T_{buf} and T_r

$$T = \beta \times T_r + (1 - \beta) \times T_{buf} \quad (6)$$

Where β is a weighting factor and its typical value is 0.5.

Our proposed improvement here: In [6], T_r is computed as follows:

$$T_r = \frac{W_{p,i}(j-1)}{W_{p,i}(j-1) \times N_{p,r} + W_{b,i}(j-1) \times N_{b,r}} \times B_i(j) \quad (7)$$

Where $W_{p,i}(j)$ and $W_{b,i}(j)$ are the complexities of previous P picture and previous B picture respectively. This weighted factor will not be accurate when the content of stream varies severely. In (5), we use ratio of the actual input bits of current picture and the remaining bits of the input GOP as the weighted factor. This factor represents the attributes of the current picture to the GOP which will be much more accurate. In case of basic unit level rate control, the sub-step 3 is to compute the target bits for each unit as follows:

$$b_l = \frac{(b_{l,org} - c_l) \times QP_{l,org}}{\sum_{k=l}^{N_{unit}} (b_{k,org} - c_k) \times QP_{l,org}} \times T_l \quad (8)$$

Where b_l is the target bits for the l^{th} unit, $b_{l,org}$, c_l and $QP_{l,org}$ are the actual bits, header bits and QP of the l^{th} unit in the input stream respectively. In this step, we use a first-order R-Q model similar to [10], in which the complexity can be computed based on bits and QP like follows:

$$X = (R - c) \times QP \quad (9)$$

Where X denotes the complexity, R denotes the total bits of the unit, and constant C is the header bits.

Step 2: After computing the target bit, QP can be computed as follows:

In case of frame level rate control:

$$Q_{step} = \frac{(T_{org} - h_{org}) \times Q_{step,org}}{T - h_{org}} \quad (10)$$

In case of basic unit level rate control:

$$Q_{step,l} = \frac{(b_{l,org} - h_{l,org}) \times Q_{step,l,org}}{b_l - h_{l,org}} \quad (11)$$

In this step, we use the header bits of input picture or basic unit as the prediction of header bits for the current picture or basic unit.

Step 3: Perform RDO for all MBs in the current basic unit and code them by H.264.

Step 4: Update the parameters

4 Experiment Results

Our experiments are based on JM7.6 codec. For simplicity, we implement a cascaded decoder-encoder transcoder with our mode-decision algorithm and rate-control algorithm. To compare the experiments results, two cascaded decoder-encoder also implemented where one of them fully decode the input stream and then re-encodes the reconstructed signals into the stream of target bit rate with RDO, and the other reuses the MB mode of input stream without RDO. We applied two types of input streams: the first one is CIF, 30Hz, 1024kbps, and

Table 2. Bit-Rate Scaling Results for Video Transcoding

Transcoding		Cascaded Decode-Encoder with RDO		Cascaded Decoder-Encoder without RDO	
Bit rate	PSNR	Bit rate	PSNR	Bit rate	PSNR
Foreman, CIF, 30Hz, 1024kbps, MB level rate control					
192.15	33.08	192.58	33.21	192.72	31.05
256.23	34.59	256.61	34.79	256.97	33.24
384.48	36.9	384.71	37.02	385.14	35.95
512.76	38.51	513	38.66	513.15	37.94
768.63	41.69	769.2	41.86	769.33	41.01

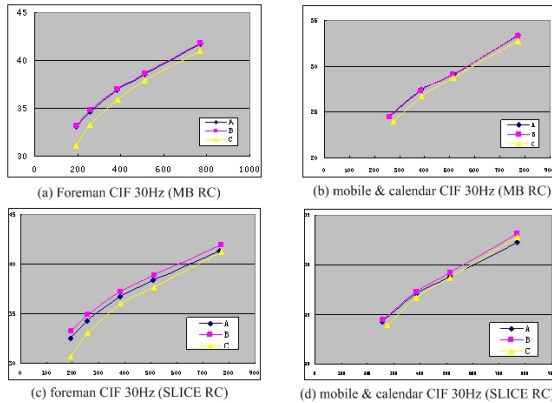


Fig. 2. R-D curves for MB and SLICE level rate control for selected sequences A: Our transcoding scheme, B: Cascaded Decoder-Encoder with RDO, C: Cascaded Decoder-Encoder without RDO

the other is QCIF, 15Hz, 384kbps. In this experiment, for simplicity, B frames are not considered. The input streams are transcoded into different bit rates.

Table 2 shows results for a set of test sequences and test conditions selected to represent a bit-rate scaling transcoding application. The result of our transcoding scheme is close to that of cascaded decoder-encoder with RDO, and re-using mode without RDO have about 1dB losses. These results also prove our rate control algorithm effective.

R-D curves for MB and SLICE level rate control for selected sequences are plotted. From fig.2 we can see that the losses of cascaded decoder-encoder without RDO increases as long as the target bit-rate decreases, while the results of our transcoding scheme performs rather well at low bit-rate.

5 Conclusion

Experimental results show that RDO is necessary in H.264 transcoding. Our fast mode decision algorithm notably reduces the complexity of RDO. Also experiments show that our rate control algorithm is simple but effective especially

when transcoding the input streams into low bit-rate streams. MB level rate control performs better than frame level and slice level rate control.

Acknowledgement. This research has been partially supported by NSFC under contract No. 60333020, “863” project under contract No. 2002AA118010, “973” project under contract No. 2001cca03300 and by Bairen project of CAS.

References

1. P. Assunao and M. Ghanbari, “Post-processing of MPEG-2 coded video for transmission at lower bit-rates,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Atlanta, CA, 1996, pp. 1998-2001.
2. Huifang Sun, Wilson Kwok and Joel W. Zdepski, “Architectures for MPEG compressed Bitstream Scaling,” IEEE Trans. CSVT, Vol. 6, No. 2, April 1996
3. Jeongnam Youn, Ming-Ting Sun, and Jun Xin, “Video Transcoder Architectures for Bit Rate Scaling of H.263 Bit Streams”, Lujun Yuan, Huifang Sun and Wen Gao, “MPEG TRANSCODING FOR DVD RECORDING”, ICICS PCM, Dec. 2003.
4. S. F. Chang and D. G. Messerschmidt, “Manipulation and composing of MC-DCT compressed video,” IEEE J. Select. Areas Commun., vol. 13, pp. 1-11, Jan. 1995
5. “Draft ITU-T recommendation and final draft standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14496-10 AVC” in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT G050, 2003
6. ISO/IEC TC JTC 1/SC 29 N5821
7. ISO/IEC 13818-2 (MPEG-2), 1994
8. ITU-T Recommendation H.263, 1995
9. “MPEG-4 Video Verification Model 18.0 (VM-18),” ISO/IEC JTC1/SC29/WG11, Doc. MPEG-N3908, 2001
10. Test Model 5, www.mpeg.org/MPEG/MSSG/tm5
11. T. Chiang and Y.-Q. Zhang, “A New Rate Control Scheme using a New Rate-Distortion Model,” IEEE Trans. CSVT, pp. 246-250, Feb. 1997
12. Mingqiang Jiang, Xiaoquan Yi, and Nam Ling, “Improved Frame-Layer Rate Control for H.264 Using MAD Ratio,”
13. Z. Li, F. Pan, K. P. Lim, G. Feng, X. Lin and S. Rahardja, “Adaptive basic unit layer rate control for JVT,” JVT-G012-r1, 7th Meeting, Pattaya II, Thailand, Mar. 2003
14. Siwei Ma, Wen Gao, Feng Wu, and Yan Lu, “Rate Control fro H.26L,”

Salient Region Detection Using Weighted Feature Maps Based on the Human Visual Attention Model

Yiqun Hu^{2,*}, Xing Xie¹, Wei-Ying Ma¹, Liang-Tien Chia², and Deepu Rajan²

¹ Microsoft Research Asia

5/F Sigma Center, No.49 Zhichun Road, P.R. China 100080

{xingx, wyma}@microsoft.com

² Center for Multimedia and Network Technology

School of Computer Engineering

Nanyang Technological University, Singapore 639798

{p030070, asltchia, asdrajan}@ntu.edu.sg

Abstract. Detection of salient regions in images is useful for object based image retrieval and browsing applications. This task can be done using methods based on the human visual attention model [1], where feature maps corresponding to color, intensity and orientation capture the corresponding salient regions. In this paper, we propose a strategy for combining the salient regions from the individual feature maps based on a new *Composite Saliency Indicator (CSI)* which measures the contribution of each feature map to saliency. The method also carries out a dynamic weighting of individual feature maps. The experiment results indicate that this combination strategy reflects the salient regions in an image more accurately.

Keywords: Salient Region Detection, Visual Attention Model, Feature Combination Strategy.

1 Introduction

In human visual system, there is a mechanism called selective attention which directs human vision to interest part(s) of visual scene. These parts are called salient regions and their saliencies correspond to how much attention can be focus on them. Visual attention analysis is generally an effective mechanism for salient region detection which is useful for region/object based image processing such as region/object based image indexing, matching, retrieval and so on. There are several computational visual attention models for simulating human visual attention [1,2,3]. Two of the most effective models are described in [1] and [2]. Both of two models use feature contrast to measure attention except for the number of feature they used. It is observed that in some cases, using only one

* This work was performed when the first author was a visiting student at Microsoft Research Asia

feature as in [2] yields salient regions that are similar to, if not better than in [1]. This phenomenon indicates that more number of features will not necessarily enhance saliency detection.

Hence, there is a need for evolving strategies to decide features that are useful and to dynamically combine them. In [4], Itti et al. compare four different feature combination strategies. Among the four strategies, the method of linear combination with learned weights is a supervised learning method which requires a prior knowledge about the salient region of the training images. Another iterative non-linear local competition strategy is proposed to overcome the defects of global non-linear normalization method. But these methods do not achieve satisfied performance across different images because all features are given positive weights even if they may erode visual attention. On the other hand, Ma and Zhang [2] consider the contrast of one fixed feature (color) for computational simplicity, but it may not be robust for the cases where color is not the most useful feature to detect saliency. A similar idea about selecting useful feature(s) for saliency is introduced in [5] where the authors select the feature map which contributes most to the strongest point in saliency map as the *winning map*. However, considering only the contribution to the strongest point cannot indicate the contribution to the whole region. Moreover, the combination of the feature maps could also result in an erroneous strongest point resulting in an erroneous selection of the feature map as the winning map.

In this paper, we present an algorithm that uses an indicator, which we call the *Composite Saliency Indicator (CSI)* to measure the contribution of each feature to the saliency map. Furthermore, we present a dynamic combination strategy to finally detect the salient regions in an image. CSI takes into account the feasibility of using a certain feature map and determines the weights to be associated with each feature map that is selected to yield the saliency map. The rest of this paper is organized as follows. In Section 2, the principle of Composite Saliency Indicator (CSI) is introduced. New feature combination strategy according to CSI is outlined in Section 3. Section 4 illustrates experiment evaluation compared with the combination methods of Itti et al. [4]. Finally the conclusion and discussion are listed in Section 5.

2 Composite Saliency Indicator

To detect salient region, we follow the model in [1] to generate the three feature maps corresponding to color, intensity and orientation. Each feature map contributes saliency differently. If strong salient points occur in a small area compared to the total size of the image, the saliency in this area can be said to be compact and distinct. An indicator called the *Composite Saliency Indicator (CSI)* is used to measure the contribution of each feature map to the salient region. The measure consists of two factors - Spatial Compactness and Saliency Density. In the following subsections we describe these in more detail.

2.1 Salient Point Selection

The first step is to detect the salient points. This is obtained by simply thresholding the color, intensity and orientation maps. The value of the threshold is decided by histogram entropy thresholding analysis [6]. Accordingly, the threshold is obtained by maximizing

$$l' = \arg \max_l \left(- \sum_{\mu=1}^l \frac{p_\mu}{\sum_{v=1}^l p_v} \log \frac{p_\mu}{\sum_{v=1}^l p_v} - \sum_{\mu=l+1}^L \frac{p_\mu}{1 - \sum_{v=1}^l p_v} \log \frac{p_\mu}{1 - \sum_{v=1}^l p_v} \right) \quad (1)$$

where p_i is the number of pixels with intensity i , L is the total number of gray levels and l is the threshold. The salient point set is defined as the set of pixels whose value is above the threshold.

2.2 Spatial Compactness

The spatial compactness of salient point set indicates the conspicuousness of potential salient region. In our work, we use convex hull to measure spatial compactness of salient point set. The procedure consists of two steps:

1. Compute convex hull polygon of salient point set using "Gift Wrapping" algorithm [7]. The algorithm begins by locating the lowest-rightmost point and then finds the point that has the smallest positive angle (with respect to the horizontal axis). A hull edge is found that joins these two points. The algorithm then proceeds to find the point with the smallest angle from this established hull edge in a counterclockwise sense. The process continues until the lowest-rightmost point is again reached.
2. Calculate the area of the polygon using the trapezoid method. We first identify the left-most vertex A of the polygon and then rearrange the vertices starting from A in a clockwise sense. The area is calculated as,

$$Size_{convexhull} = \sum_{k=1}^{K+1} \frac{(y_{k+1} + y_k) \times (x_{k+1} - x_k)}{2} \quad (2)$$

where (x_i, y_i) are co-ordinates of the vertex and K is the total number of vertices in the convex hull.

Figure 1 is an example of convex hull calculation. For the images with multiple salient regions, the above procedure can be applied to each connected salient point subset and summed up over all convex hulls to get a measure of spatial compactness.

2.3 Saliency Density

The second part of the CSI is the saliency density indicated by the gray-level value of the points. Spatial compactness only considers the spatial relationship

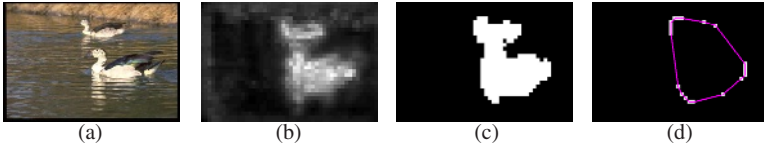


Fig. 1. Example of Spatial Compactness Measure using Convex Hull; (a) Original Image; (b) Intensity Feature Map; (c) Salient Point Set; (d) Convex Hull

of salient points. Two salient point sets with the same spatial compactness can have different effects for salient region indication. The feature map with strong saliency density in one or few specific areas is better for extracting salient region. We calculate saliency density as

$$D_{saliency} = \frac{\sum_{p \in \theta} \frac{\sum_{q \in \theta_n(p)} |I(p) - I(q)|}{|\theta_n(p)|}}{|\theta|} \quad (3)$$

Where $I(x)$ is the intensity at location x , $\theta_n(p)$ is the set of all neighboring salient points of p and θ is the set of salient points. If all salient points with similar saliency value are close to each other, $D_{saliency}$ will give a small value indicating that the saliency of this map is conspicuous. A large value of $D_{saliency}$ implies inconspicuity of saliency. Note that the saliency density is measured using intensity values from the feature map but only for those locations in the salient point set.

3 Feature Combination Strategy

Based on the saliency measure $Size_{convexhull}$ and $D_{saliency}$, a two level feature combination strategy is designed for feature map combination. The advantage of the proposed combination strategy lies in that it dynamically decides whether a feature will be selected, and if so, what will be its weight.

The feature maps are first classified into two categories according to its spatial compactness. If $Size_{convexhull} < 80\%$ of the feature map area, we call it a *Non-uniform Map*, else it is called a *Uniform Map*. In a uniform map, the saliency is not sparse implying that it does not contain much useful information with regard to human visual attention; hence, a uniform map is not considered while determining the saliency map. Among non-uniform feature maps, the feature map with smallest $Size_{convexhull}$ is selected as the *Reference Map* (RM). Then the similarity between each non-uniform map and the *Reference Map* is examined. The similarity measure used is

$$d_p(f_i, f_j) = \left(\sum_{m=1}^M \sum_{n=1}^N |f_i(m, n) - f_j(m, n)|^p \right)^{1/p} \quad (4)$$

where f_i and f_j are the $M \times N$ feature maps and $p = 2$ for Euclidean distance. The projection vectors are used to reduce the dimension of the feature space.

Table 1. Rules for Feature Map Combination

Index	Concept	Condition	Combination Strategy
1	Uniform Map	$Size_{convexhull} > 80\%$ of map area	Skip during combination
2	Reference Map	$\min Size_{convexhull}$	Weighted combination
3	Related Map	$\hat{d}_p(f_i, f_{RM}) < T_t$	Weighted combination
4	Unrelated Map	$\hat{d}_p(f_i, f_{RM}) \geq T_t$	Skip during combination

They are denoted by l_n^r and l_m^c for the n^{th} row and m^{th} column, respectively, i.e., $l_n^r(f) = \sum_{m=1}^M f(m, n)$ and $l_m^c(f) = \sum_{n=1}^N f(m, n)$. Substituting these in equation (4), we get

$$\hat{d}_p(f_i, f_j) = \left(\sum_{n=1}^N \left| \frac{1}{M} (l_n^r(f_i) - l_n^r(f_j)) \right|^p + \sum_{m=1}^M \left| \frac{1}{N} (l_m^c(f_i) - l_m^c(f_j)) \right|^p \right)^{\frac{1}{p}} \quad (5)$$

Based on this similarity measure, non-uniform maps are further divided into two categories - if $\hat{d}_p(f_i, f_{RM}) < T_t$, where T_t is a threshold, then we call the non-uniform map as *Related Map*, else it is called an *Unrelated Map*. Unrelated maps are ignored in the combination. In the second level, the related maps are linearly combined with the reference map. The weighting coefficients are calculated according to spatial compactness and saliency density according to

$$W_{total} = \sum_{p \in \Phi} (Size_{convexhull}(p) \times D_{saliency}(p)) \quad (6)$$

$$W_i = \frac{Size_{convexhull}(i) \times D_{saliency}(i)}{\sqrt{\sum_{p \in \Phi} (Size_{convexhull}(p) \times D_{saliency}(p))^2}} \quad (7)$$

where ϕ is the set of all *Related Maps*.

Table 1 summarizes the rules of the proposed feature combination strategy. Figure 2 compares the result of the combination strategy proposed in this paper with that in [1]. Since the salient point sets corresponding to the orientation map and the intensity map do not satisfy the saliency rules, they are ignored. As a result, We see that the detected salient region is more compact and closer to the human visual system using the method described in this paper. The dynamic weighting scheme of the proposed strategy is illustrated in Figure 3. Notice that the weights are chosen according to the *Related Map* and the *Reference Map* differently in different images. After combining all related feature maps and generating a global saliency map, any region extraction methods can be used to extract salient region such as Seeded Region Growing [8].

4 Experiments and Evaluation

600 images are randomly selected from the standard Corel Photo Library as the data set to evaluate the performance of the proposed method. Figure 4

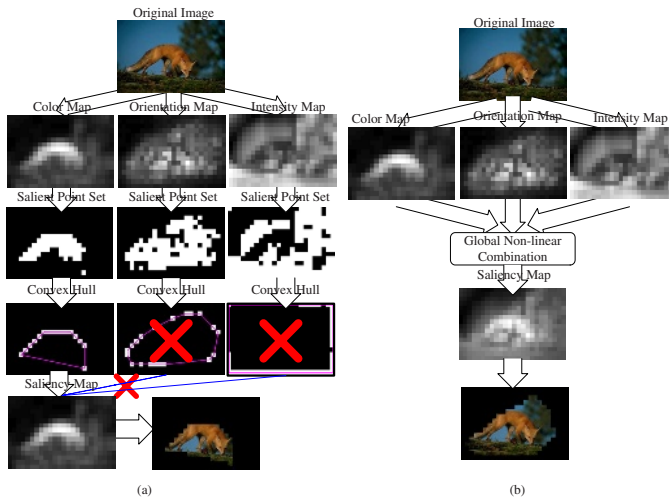


Fig. 2. Comparison of (a) the proposed combination strategy using CSI with that of (b) non-linear combination [1].

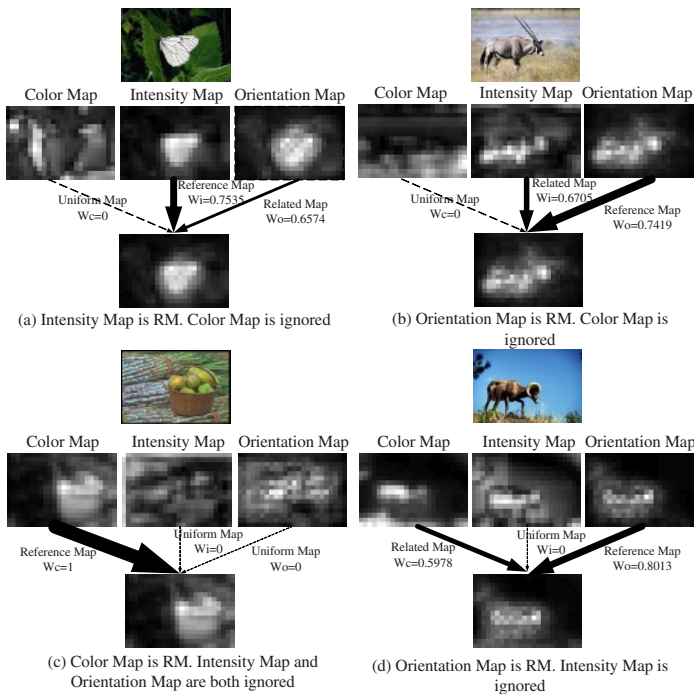


Fig. 3. Dynamic Combination Strategy

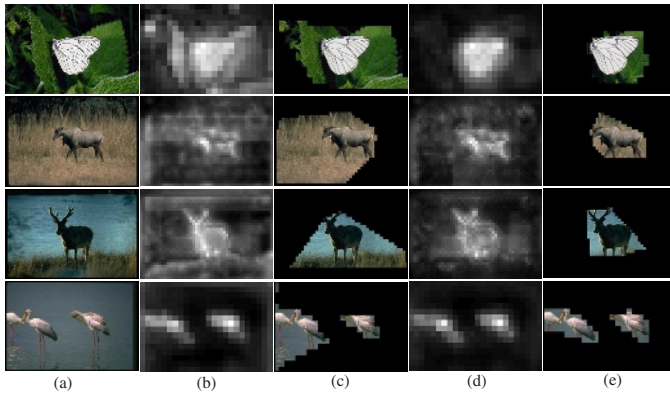


Fig. 4. (a) Original Image; (b) Saliency Map using Itti et al's model [4]; (c) Cropped Image using Itti et al's model [4]; (d) Saliency Map using CSI; (e) Cropped Image using CSI.

Table 2. User Study Result Evaluation

User	CSI Better	Non-linear combination Better	Both Equally Good
1	55.0%	10.0%	35.0%
2	57.5%	10.0%	32.5%
3	45.0%	17.5%	37.5%
4	57.5%	7.5%	35.0%
5	57.5%	10.0%	32.5%
6	52.5%	7.5%	40.0%
Average	54.2%	10.4%	35.4%

shows several examples of the experiment. Notice that the saliency map obtained using the proposed CSI reflects the salient regions more accurately than that obtained by the non-linear combination method of [1]. Our method is also able to successfully capture more than one salient region as shown in the last row of Figure 4. Due to the subjective nature of the problem, a user study was conducted to evaluate the results of the experiment.

Six subjects are invited to each view any 40 of the 600 images. The subjects were asked if the cropped regions reflected the human visual attention region of the image for the proposed method as well as for the the method of [1]. Table 2 shows the result of the user study. The proposed combination strategy using CSI outperforms the non-linear combination strategy in more than 50% of the cases. About 35% of the responses indicate that both the strategies are equally good. However, about 10% of the responses suggest that the output of the non-linear combination strategy was better. This can be attributed to the incorrect threshold values selected from the entropy thresholding model discussed in Section 2.1. We point out that if any one of the feature maps shows a salient region clearly, we get a better result of the cropped image using CSI. However, if none

of the feature maps shows a distinct salient region, the result is no worse than the non-linear combination strategy.

5 Conclusion and Discussion

In this paper, we propose a method to identify useful feature maps that capture salient regions based on the human visual attention model as well as a method to dynamically weight each of the selected feature maps to locate salient regions in an image. Compared to existing feature combination strategies, it improves the accuracy of salient region detection. The improvement of proposed feature combination strategy according to the new proposed measure of CSI is useful in subsequent processing such as object extraction. Such object extraction methods can then be used for image retrieval and browsing [9]. Further extensive subjective test and salient region detection in clutter scene will be investigated in our future work.

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (1998) 1254–1259
2. Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. In: *Proceedings of the eleventh ACM international conference on Multimedia*. Volume 1. (2003) 374–381
3. Kadir, T., Brady, M.: Saliency, scale and image description. *International Journal of Computer Vision* **45** (2001) 83–105
4. Itti, L., Koch, C.: A comparison of feature combination strategies for saliency-based visual attention systems. In: *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, San Jose, CA. Volume 3644. (1999) 473–482
5. Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C.: Attentional selection for object recognition - a gentle way. *Lecture Notes in Computer Science* **25** (2002) 472–279
6. Wong, A., Sahoo, P.: A gray-level threshold selection method based on maximum entropy principle. *IEEE Transactions on Systems, Man, and Cybernetics* (1989) 866–871
7. Sugihara, K.: Robust gift wrapping for the three-dimensional convex hull. *J. Comput. Syst. Sci.* **49** (1994) 391–407
8. Adams, R., Bischof, L.: Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16** (1994) 641–647
9. Hu, Y., Xie, X., Ma, W.Y., Rajan, D., Chia, L.T.: Salient object extraction combining visual attention and edge information. Technical Report (2004)

An Attention-Based Decision Fusion Scheme for Multimedia Information Retrieval

Xian-Sheng Hua and Hong-Jiang Zhang

Microsoft Research Asia
3F Sigma Center, 49 Zhichun Road, BEIJING 100080, P.R.CHINA
{xshua,hjzhang}@microsoft.com

Abstract. In this paper, we proposed a novel decision fusion scheme based on the psychological observations on human beings' visual and aural attention characteristics, which combines a set of decisions obtained from different data sources or features to generate better decision result. Based on studying of the "heterogeneity" and "monotonicity" properties of certain types of decision fusion issues, a set of so-called *Attention Fusion Functions* are devised, which are able to obtain more reasonable fusion results than typical fusion schemes. Preliminary experiment on image retrieval shows the effectiveness of the proposed fusion scheme.

1 Introduction

Generally information fusion is about summarizing information in an embodiment of multiple sources and typically categorized into three classes, data fusion, feature fusion and decision fusion [1]. Considerable works on this topic have been reported in literatures [1]-[8]. In the field of multimedia content analysis, indexing and retrieval, numerous theory or application issues can be classified into one or more of these three classes. Among these, decision fusion is aiming at obtaining better or optimal decision result by appropriately combining a set of decision results, whether they are hard decisions (i.e., true or false) or soft ones (confidence values), from different sensors, feature sets, and so on. A good decision fusion scheme is expected to sufficiently utilize the information provided by the set of to-be-fused decision results, while suppress the noises in them at the same time.

In this paper, we will propose a new decision fusion scheme based on the psychological observations on human being's visual and aural attention characteristics, which combines a set of decisions obtained from different data sources or features to generate better decision result. "Attention" is a neurobiological conception. It means the concentration of mental powers upon an object by close or careful observing or listening, which is the ability or power to concentrate mentally [9]. Generally, people will concentrate their attention upon the circumstances when there is something (event, object, etc.) can be apperceived. And the "degree" that people will concentrate their attention on the event or object is proportional to the "quantity" of the information (or we may call it

attention index or *attention value* [9]) it provides, say, the strength of a sound, the speed of a motion, the size of an object, and so on. But what will happen when two or more sources of such information are provided in the circumstance? Are people will concentrate their attention in the degree of the “sum” of all of the attention indices, or the average of them, or the maximum of them? Generally linear combination (*LC*) of the attention indices of different attention components (e.g., motion, color, audio, etc.) is a simple but effective scheme. However, this kind of linear combination is not reflecting all the information that the attention indices of the attention components contained [10]. In this paper, we proposed a so-called *Attention Fusion Function (AFF)* to model the above issues. And, this fusion scheme can be used to fuse a set of decisions with similar characteristics (we may call it “attention properties”, to be detailed in Section 2). It should be mentioned that the proposed *AFF* fusion scheme is not a general fusion scheme suited for solving general decision fusion issues, but especially applicable for decision issues having the “attention properties”.

The rest of the paper is organized as follows. Section 2 presents the attention-based decision fusion scheme in detail. Preliminary experiments on applying the fusion scheme on image retrieval are introduced in Section 3, followed by conclusion remarks in Section 4.

2 Attention Fusion

We denote the set of to-be-fused decision results (normalized to interval $[0,1]$) as a decision vector $\vec{x} = (x_1, x_2, \dots, x_n)$, where $0 \leq x_i \leq 1, 1 \leq i \leq n$. A general fusion function is denoted as $f(\vec{x})$ or $f(x_1, x_2, \dots, x_n)$. It should be noted that the following analyses are based on the assumption that the to-be-fused decisions have the “attention properties” mentioned above.

2.1 Two-Dimensional Case

Firstly let’s consider a simple case in which we only have two decisions to fuse, i.e., $n=2$. Weights for the decisions are also not taken into account at this stage, i.e., all the weights are equal to $1/n$. Let’s see two decision vectors, $(0.8, 0)$ and $(0.4, 0.4)$, if linear combination is applied, $f(0.8, 0)$ will have the same value as $f(0.4, 0.4)$. However, this result does not coincide with the real case. Actually, the first decision vector is more “attractive” (if we regard the decisions as attention indices of different attention components), as one attention component with high attention index will “attract” people’s attention greatly. Accordingly, it will be better if the fusion function satisfies the following inequality,

$$f(x_1, x_2) < f(x_1 + \varepsilon, x_2 - \varepsilon), \quad (1)$$

where $0 < \varepsilon \leq x_2 \leq x_1$. In contrast, if we use linear combination, $f(x_1, x_2)$ is equal to $f(x_1 + \varepsilon, x_2 - \varepsilon)$, thus it does not satisfy this property. For convenience,

we name this property “heterogeneity”. On the other hand, it is obvious that the fusion function should satisfy another property, monotonicity, i.e.,

$$f(x_1, f_2) < f(x_1 + \varepsilon, x_2), \tag{2}$$

where $\varepsilon > 0$. Maximum function (*MAX*) satisfies (1), while does not strictly satisfy this monotonicity property. When the strict inequality signs in equality (1) and (2) are replaced by non-strict inequality signs, the fusion function could be,

$$AFF^{(0)} = \frac{1}{2} [(x_1 + x_2) + |x_1 - x_2|]. \tag{3}$$

Obviously, this function is obtained just by adding a correction, the difference between the two decisions, to the linear combination fusion result. In order to strictly satisfy the two inequalities simultaneously, we have the following theorem.

Theorem 1. (*2-Dimensional AFF without Weights*): *The following function*

$$AFF_2^{(\gamma)}(x_1, x_2) = \frac{1}{2} [(x_1 + x_2) + \frac{1}{1 + \gamma} |x_1 - x_2|]. \tag{4}$$

satisfies inequality (1) and (2), where $\gamma > 0$ is a constant. □

For the above mentioned example, we got $AFF_2^{(0.2)}(0.8, 0) = 0.733$, while $AFF_2^{(0.2)}(0.4, 0.4) = 0.4$, which obviously indicates the first one is more “attractive”. Fig. 1 shows the comparison of *LC*, *MAX*, $AFF^{(0)}$ and $AFF_2^{(\gamma)}$. Actually if $x_1 = x_2$, we will have $MC = MAX = AFF_2^{(\gamma)} = AFF^{(0)}$, while when $x_1 \neq x_2$, we have

$$LC(x_1, x_2) < AFF_2^{(\gamma)} < AFF_2^{(0)}(x_1, x_2) = MAX(x_1, x_2). \tag{5}$$

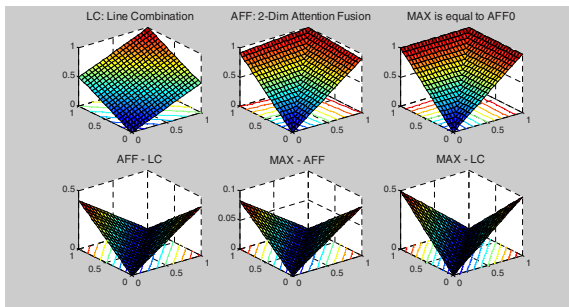


Fig. 1. Comparison of *LC*, *MAX*, $AFF^{(0)}$ (denoted by “AFF0” in the figure), and $AFF_2^{(\gamma)}$ (denoted by “AFF”). The three sub-figures in the second line show the differences between $AFF_2^{(\gamma)}$ and *LC*, *MAX* and $AFF_2^{(\gamma)}$, *MAX* and *LC*, respectively.

Table 1. Differences between attention fusion function and averaging.

x_1	x_2	γ	$AF F_2^{(\gamma)}$	Average	Difference	γ	$AF F_2^{(\gamma)}$	Average	Difference
0.1	0.8		0.74	0.45	0.29		0.68	0.45	0.23
0.2	0.8		0.75	0.50	0.25		0.70	0.50	0.20
0.5	0.5	0.2	0.50	0.50	0.00	0.5	0.50	0.50	0.00
0.0	1.0		0.91	0.50	0.41		0.83	0.50	0.33
0.4	0.6		0.67	0.50	0.17		0.63	0.50	0.13

The parameter $\gamma > 0$ is a predefined constant, which controls the amount of differences between the left sides and right sides of inequalities (1) and (2) when x_1, x_2 and ε are fixed. The greater the parameter ε is, the smaller the differences are. To be exact, $(1/\gamma)$ represents the effectiveness of one decision component in the overall decision. For example, $[f(0.5, 0.7) - f(0.6, 0.6)]$ is equal to 0.091 and 0.067 when γ is equal to 0.1 and 0.5, respectively. The smaller the parameter γ is, the more greatly that one decision component with high index (or confidence) will affect (increase) the overall decision index. In Table 1, we list some examples to show the differences between attention fusion function and LC (i.e., average) under different parameters. In fact, there may be other fusion functions also satisfy both heterogeneity and monotonicity; however, the proposed one in this paper actually is a set of functions controlled by parameter γ , which may be appropriately adjusted to fit different requirements for different applications.

2.2 n -Dimensional Case

For n -dimensional case we have Theorem 2.

Theorem 2. (*n -Dimensional AFF without Weights*): The following function

$$AF F_n^{(\gamma)}(\vec{x}) = E(\vec{x}) + \frac{1}{2(n-1) + n\gamma} \sum_{k=1}^n |x_k - E(\vec{x})|, \tag{6}$$

where $\gamma > 0$ is a predefined constant, and $E(\vec{x})$ is the average value of decision components in vector \vec{x} , satisfies the inequality

$$AF F_n^{(\gamma)}(x_1, \dots, x_i, \dots, x_n) < AF F_n^{(\gamma)}(x_1, \dots, x_i + \varepsilon, \dots, x_n), \tag{7}$$

where $1 \leq i \leq n$, $\varepsilon \geq 0$, and inequality

$$AF F_n^{(\gamma)}(x_1, \dots, x_i, \dots, x_j, \dots, x_n) < AF F_n^{(\gamma)}(x_1, \dots, x_i + \varepsilon, \dots, x_j - \varepsilon, x_n), \tag{8}$$

where $1 \leq i < j \leq n$, $x_i \geq x_j \geq \varepsilon > 0$. When

$$x_j - \varepsilon \leq E(\vec{x}) \leq x_i + \varepsilon, \tag{9}$$

the inequality sign in (8) will be strictly satisfied. □

When weights are taken into consideration, we have Theorem 3.

Theorem 3. (*n-Dimensional AFF with Weights*): The following function

$$AFFW_n^{(\gamma)}(\vec{x}) = \frac{\vec{w} \cdot \vec{x} + \frac{1}{2(n-1)+n\gamma} \sum_{k=1}^n |nw_k x_k - \vec{w} \cdot \vec{x}|}{W}, \tag{10}$$

satisfies inequality

$$AFFW_n^{(\gamma)}(x_1, \dots, x_i, \dots, x_n) < AFFW_n^{(\gamma)}(x_1, \dots, x_i + \varepsilon, \dots, x_n), \tag{11}$$

where $1 \leq i \leq n$, $\varepsilon \geq 0$, and inequality

$$AFFW_n^{(\gamma)}(x_1, \dots, x_i, \dots, x_j, \dots, x_n) < AFFW_n^{(\gamma)}(x_1, \dots, x_i + w_j \varepsilon, \dots, x_j - w_i \varepsilon, \dots, x_n), \tag{12}$$

where $1 \leq i < j \leq n$, $w_i x_i \geq w_j x_j \geq w_i w_j \varepsilon > 0$. When

$$nw_j x_j - nw_i w_j \varepsilon \leq \vec{w} \cdot \vec{x} \leq nw_i x_i + nw_i w_j \varepsilon, \tag{13}$$

the inequality sign in (12) will be strictly satisfied. In the above expressions, $\gamma > 0$ is a predefined constant, and

$$W = W(\vec{w}) = 1 + \frac{1}{2(n-1) + n\gamma} \sum_{k=1}^n |1 - nw_k|, \tag{14}$$

and $\vec{w} = (w_1, w_2, \dots, w_n)$ is the weight vector satisfying $w_k \geq 0$, $1 \leq k \leq n$, and $\sum_{k=1}^n w_k = 1$. \square

Actually, W is equal to the maximum value of the numerator of the right side of equation (10). Therefore, we have

$$0 \leq AFFW_n^{(\gamma)}(\vec{x}) \leq 1. \tag{15}$$

When all weights are equal, obviously we have $W = 1$ and

$$AFFW_n^{(\gamma)}(\vec{x}) = AFF_n^{(\gamma)}(\vec{x}). \tag{16}$$

Fig. 2 shows the comparison of LC , $AFF_2^{(\gamma)}$ and $AFFW_2^{(\gamma)}$, when $w_1 = 0.6$ and $w_2 = 0.4$.

2.3 Further Improvement

It is observed that, for n -dimensional AFF , some fusion results are still not “fair” enough. For example, when $n = 10$, $f(0.1, 0.1, \dots, 0.1) = 0.1$ while $f(1, 0, \dots, 0) = 0.19$ ($\gamma = 0.2$). Intuitively, $f(1, 0, \dots, 0)$ should have relatively larger value. Actually, according to (6) and (8), it is easy to prove that

$$LC(\vec{x}) \leq AFF_n^{(\gamma)}(\vec{x}) \leq AFF_n^{(0)}(\vec{x}) \leq \frac{2MAX(\vec{x})}{n}. \tag{17}$$

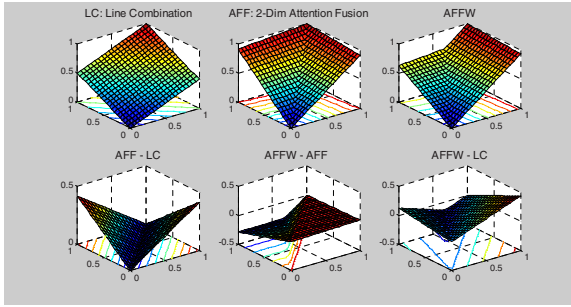


Fig. 2. Comparison of $LC, AFF_2^{(\gamma)}$ (denoted by “AFF” in the figure), and $AFFW_2^{(\gamma)}$ (denoted by “AFFW”). The figures in the second line show the differences between $AFF_2^{(\gamma)}$ and $LC, AFFW_2^{(\gamma)}$ and $AFF_2^{(\gamma)}$, $AFFW_2^{(\gamma)}$ and LC , respectively.

Therefore, when n is large, $AFF_n^{(\gamma)}(\vec{x})$ is much less than $MAX(\vec{x})$. To solve this issue, another constraint is required, which is defined by

$$f(\vec{x}) \geq \alpha \cdot MAX(\vec{x}), \tag{18}$$

where α is a predefined constant.

Theorem 4. (Improved n -Dimensional AFF with Weights): The following function

$$IAFFW_n^{(\gamma)}(\vec{x}) = \frac{\alpha \cdot \max_{1 \leq k \leq n} \{nw_k x_k\} + \beta \cdot W \cdot AFFW_n^{(\gamma)}(\vec{x})}{W^*} \tag{19}$$

satisfies inequality (11), (12) (function name should be replaced by $IAFFW$) under the corresponding constraints, as well as satisfies

$$IAFFW_n^{(\gamma)}(\vec{x}) \geq \frac{\alpha}{W^*} \cdot \max_{(1 \leq k \leq n)} \{nw_k x_k\}, \tag{20}$$

$$0 \leq IAFFW_n^{(\gamma)}(\vec{x}) \leq 1, \tag{21}$$

where $0 < \alpha, \beta < 1$, $\alpha + \beta = 1$, W is defined in equation (14), and

$$W^* = \alpha \cdot \max_{1 \leq k \leq n} \{nw_k x_k\} + \beta \cdot W. \tag{22}$$

When we have (13) or

$$w_i x_i = \max_{1 \leq k \leq n} \{w_k x_k\}, \tag{23}$$

the inequality sign in (12) will be strictly satisfied. \square

Proof of Theorem 4 can be found in the Appendix, which also indicates Theorem 1, 2 and 3 are also correct. Fig. 3 shows the comparison of $AFFW_2^{(0.2)}(\vec{x})$ and $IAFFW_2^{(0.2)}(\vec{x})$. In the above example, when we set

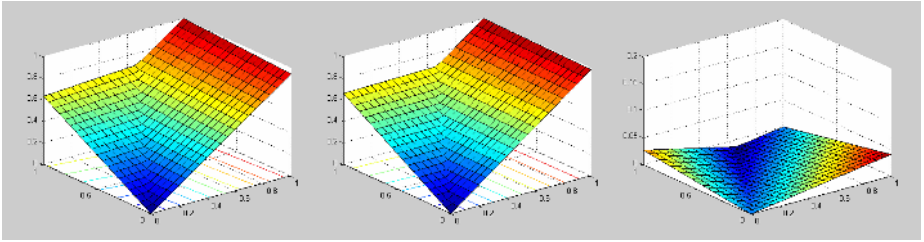


Fig. 3. Comparison of $IAFW_2^{(0.2)}$ and $AFFW_2^{(0.2)}$ ($\alpha = 0.7, \beta = 0.3, w_1 = 0.6, w_2 = 0.4$)

Table 2. Comparison of different decision fusion schemes. **R100** is the recall rate for the first 100 query results, while **P10, P20, P30** and **P100** are the precision rates for the first 10, 20, 30 and 100 query results, respectively. And the parameters for **IAFFW** are set as, $\gamma = 0.2, \alpha = 0.7, \beta = 0.3, w_1 = w_2 = \dots = w_6 = 1/6$.

Fusion Schemes	R100	P10	P20	P30	P100
Average Fusion	0.121	0.315	0.244	0.212	0.138
Maximal Fusion	0.125	0.321	0.252	0.230	0.141
IAFFW	0.130	0.325	0.265	0.243	0.153

$\alpha = 0.7, \beta = 0.3, \gamma = 0.2$, and all weights are equal to 0.1, we have better results as $f(0.1, 0.1, \dots, 0.1) = 0.1$ and $f(1, 0, \dots, 0) = 0.757$. Obviously, if all weights are equal, we have $W^* = W = 1$, and the improved attention fusion function without weight is

$$IAFW_n^{(\gamma)}(\vec{x}) = \alpha \cdot \max_{1 \leq k \leq n} \{x_k\} + \beta \cdot AFF_n^{(\gamma)}(\vec{x}). \tag{24}$$

3 Preliminary Experiments

Content-based image retrieval is taken as an example to demonstrate the advantage of the novel fusion scheme. A database containing 10,000 images excerpted from Coral Draw image database is used, which has been categorized into 79 classes (such as tiger, beach, building, etc.) according to their content. In the experiments, 20% images randomly selected from the database are applied as query samples one-by-one, while the images in the same class as the query image are considered as positive samples.

Six sets of features, including color histogram, color moment, wavelet, block wavelet, correlogram, blocked correlogram [11], are employed to calculate the similarity between two images using $L1$ distance measure. Each set of features will produce a similarity measure or decision for any image in the database compared with the query image. Table 2 shows the performances (recall rate for the first 100 query results and precision rate for the first 10, 20, 30 and

100 query results) when using average fusion, maximal fusion and attention fusion (*IAFFW*), respectively. From the numbers in Table 2, it can be seen that Attention Fusion produces relatively better results.

A better Evaluation of the proposed attention fusion scheme can be obtained by applying the scheme on the decisions produced from different types of features, such as color, shape and so on, as well as by testing on applications in which the decisions are obtained from multiple modalities. This is our future work.

4 Conclusion and Discussion

In this paper, we have proposed a novel decision fusion scheme, Attention Fusion, which combines a set of decisions generated from different data sources or features to obtain better decision result. The fusion scheme is based on the two properties, monotonicity and heterogeneity, of the to-be-fused decision set, which come from psychological observations and assumptions of human beings' visual and aural attentions. The proposed fusion scheme can be used to obtain better decision result in the case of these two properties are satisfied. The future work would be to test the scheme on more and wider applications. In addition, how to automatically determine the best parameters is still unsolved. Another future work would be to extend the idea of constructing attention functions to other types of decision fusion issues with different properties.

References

1. Dasarathy, B.V., Decision Fusion. IEEE Computer Society, August 1993.
2. Samarasooriya, V.N.S., et al, A Fuzzy Modeling Approach to Decision Fusion Under Uncertainty. IEEE /SICE/RSJ Intl. Conf. on Multisensor Fusion and Integration of Intelligent Systems (1996), 788-795.
3. Chen, B., Varshney, P.K., A Bayesian Sampling Approach to Decision Fusion Using Hierarchical Models. IEEE Trans. on Signal Processing (2002), Vol. 50, No. 8.
4. Li, X.R., Zhu, Y., Wang, J., Han, C., Optimal Linear Estimation Fusion - Part I: Unified Fusion Rules. IEEE Trans. on Information Theory (2003), Vol. 49, No. 9.
5. Woods, K., et al, Combination of Multiple Classifiers Using Local Accuracy Estimates, IEEE Transactions on Pattern Analysis and Machine Intelligence (1997), April.
6. Petrakos, M., et al, The Effect of Classifier Agreement on the Accuracy of the Combined Classifier in Decision Level Fusion, IEEE Transactions on Geoscience and Remote Sensing (2001), Nov.
7. Ji, C., Ma, S., Combinations of Weak Classifiers. IEEE Transactions on Neural Networks (1997), Jan.
8. Xu, L., et al, Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition, IEEE Transactions on Systems, Man and Cybernetics(1992), May/June.
9. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.J., An Attention Model for Video Summarization. ACM Multimedia (2002), Juan-les-Pins, France, December.

10. Ma, Y.F., et al, User Attention Model based Video Summarization. To appear in IEEE Transactions on Multimedia Journal.
11. Veltkamp, R. C., Tanase, M., Content-based image retrieval systems: a survey. March 2001, <http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/>.

Appendix

Proof of Theorem 4: Without losing generality, we let $i = 1$, and $j = 2$.

(a) Proof of inequality (12):

$$\begin{aligned}
 IAFFW_n^{(\gamma)}(\vec{x}') &= IAFFW_n^{(\gamma)}(x'_1, x'_2, \dots, x'_n) \\
 &= IAFFW_n^{(\gamma)}(x_1 + w_2\varepsilon, x_2 - w_1\varepsilon, x_3, \dots, x_n) \\
 &= \frac{\alpha \cdot \max\{nw_1x_1 + nw_1w_2\varepsilon, nw_2x_2 - nw_1w_2\varepsilon, nw_3x_3, \dots, nw_nx_n\}}{W^*} \\
 &\quad + \frac{\beta \cdot W \cdot AFW_n^{(\gamma)}(\vec{x}')}{W^*}.
 \end{aligned} \tag{25}$$

As $w_1x_1 \geq w_2x_2 \geq w_1w_2\varepsilon$, we have

$$\max\{nw_1x_1 + nw_1w_2\varepsilon, nw_2x_2 - nw_1w_2\varepsilon, nw_3x_3, \dots, nw_nx_n\} \geq \max_{1 \leq k \leq n} \{nw_kx_k\}. \tag{26}$$

On the other hand,

$$W \cdot AFW_n^{(\gamma)}(\vec{x}) = \frac{1}{n} \sum_{k=1}^n nw_kx_k + \frac{1}{2(n-1) + n\gamma} \sum_{k=1}^n |nw_kx_k - \frac{1}{n} \sum_{k=1}^n nw_kx_k|. \tag{27}$$

Let

$$y_k = nw_kx_k, y'_k = nw_kx'_k, \varepsilon' = nw_1w_2\varepsilon. \tag{28}$$

Then we have

$$W \cdot AFW_n^{(\gamma)}(\vec{x}) = E(\vec{y}) + \frac{1}{2(n-1) + n\gamma} \sum_{k=1}^n |y_k - E(\vec{y})| \tag{29}$$

and

$$W \cdot AFW_n^{(\gamma)}(\vec{x}') = E(\vec{y}') + \frac{1}{2(n-1) + n\gamma} \sum_{k=1}^n |y'_k - E(\vec{y}')|. \tag{30}$$

As

$$y_1 = nw_1x_1 \geq nw_2x_2 \geq y_2, \tag{31}$$

it is easy to prove (by removing the absolute signs under certain conditions) that

$$\begin{aligned}
 |y'_1 - E(\vec{y}')| + |y'_2 - E(\vec{y}')| &= |y_1 + \varepsilon' - E(\vec{y}')| + |y_2 - \varepsilon' - E(\vec{y}')| \\
 &\geq |y_1 - E(\vec{y}')| + |y_2 - E(\vec{y}')|.
 \end{aligned} \tag{32}$$

From (29), (30) and (32), we have

$$W \cdot AFFW_n^{(\gamma)}(\vec{x}') \geq W \cdot AFFW_n^{(\gamma)}(\vec{x}). \tag{33}$$

Consequently, from (25), (26) and (33) we have

$$IAFFW_n^{(\gamma)}(\vec{x}') \geq IAFFW_n^{(\gamma)}(\vec{x}). \tag{34}$$

When (13) is satisfied, the inequality sign in (32) and (33) will be strictly satisfied. While if (23) is satisfied, the inequality sign in (26) will be strictly satisfied. Therefore, when (13) or (23) is satisfied, the inequality sign in (34) will strictly satisfied.

(b) Proof of inequality (11): Similar to (a), we only need to verify

$$W \cdot AFFW_n^{(\gamma)}(\vec{x}') > W \cdot AFFW_n^{(\gamma)}(\vec{x}), \tag{35}$$

where

$$\vec{x}' = (x_1 + \varepsilon, x_2, \dots, x_n). \tag{36}$$

We use (28) to define $\vec{y} = (y_1, y_2, \dots, y_n)$ and $\vec{y}' = (y'_1, y'_2, \dots, y'_n)$, we then have

$$\begin{aligned} & [2(n-1) + n\gamma]E(\vec{y}') + \sum_{k=1}^n |y'_k - E(\vec{y}')| \\ &= [2(n-1) + n\gamma](E(\vec{y}) + \frac{w_1\varepsilon}{n}) + \\ & \quad |y_1 + w_1\varepsilon - (E(\vec{y}) + \frac{w_1\varepsilon}{n})| + \sum_{k=2}^n |y_k - (E(\vec{y}) + \frac{w_1\varepsilon}{n})| \end{aligned} \tag{37}$$

$$\begin{aligned} &= [2(n-1) + n\gamma]E(\vec{y}) + [|y_1 + w_1\varepsilon - (E(\vec{y}) + \frac{w_1\varepsilon}{n})| + \frac{n-1}{n}w_1\varepsilon] \\ & \quad + \sum_{k=2}^n [|y_k - (E(\vec{y}) + \frac{w_1\varepsilon}{n})| + \frac{w_1\varepsilon}{n}] + \gamma w_1\varepsilon \end{aligned} \tag{38}$$

$$\geq [2(n-1) + n\gamma](E(\vec{y}) + |y_1 - E(\vec{y})|) + \sum_{k=2}^n |y_k - E(\vec{y})| + \gamma w_1\varepsilon \tag{39}$$

$$> [2(n-1) + n\gamma](E(\vec{y}) + \sum_{k=1}^n |y_k - E(\vec{y})|). \tag{40}$$

Thus (35) is proved, and consequently inequality (11) is proved.

(c) Proofs of inequality (20) and (21) are obvious, so they are omitted here. □

Approximating Inference on Complex Motion Models Using Multi-model Particle Filter

Jianyu Wang¹, Debin Zhao^{1,2}, Shiguang Shan¹, and Wen Gao^{1,2}

¹ Department of Computer Science, Harbin Institute of Technology, China

² JDL, Institute of Computing Technology, China Academy of Sciences
{jywang,dbzhao,sgshan,wgao}@jd1.ac.cn

Abstract. Due to its great ability of conquering clutters, which is especially useful for high-dimensional tracking problems, particle filter becomes popular in the visual tracking community. One remained difficulty of applying the particle filter to high-dimensional tracking problems is how to propagate particles efficiently considering complex motions of the target. In this paper, we propose the idea of approximating the complex motion model using a set of simple motion models to deal with the tracking problems cumbered by complex motions. Then, we provide a practical way to do inference on the set of simple motion models instead of original complex motion model in the particle filter. This new variation of particle filter is termed as Multi-Model Particle Filter (MMPF). We apply our proposed MMPF to the problem of head motion tracking. Note that the defined head motions include both rigid motions and non-rigid motions. Experiments show that, when compared with the standard particle filter, the MMPF works well for this high-dimensional tracking problem with reasonable computational cost. In addition, the MMPF may provide a possible solution to other high-dimensional sequential state estimation problems such as human body pose estimation and sign language estimation and recognition from video.

1 Introduction

Many researchers make extensive efforts in the visual tracking area and two decades of research have yielded many powerful tracking systems [1,3,4,5,6,8, e.g.]. One remained challenging problem for visual tracking is how to deal with tracking problems cumbered by high-dimensional complex motions robustly and efficiently, such as non-rigid head motion estimation, hand pose and body pose recognition.

Due to doing inference under the Bayesian framework and not assuming the distribution form of the posterior, particle filter, also known as CONDENSATION in the computer vision community [1,4], becomes popular and is one of the promising techniques to deal with complex tracking problems with the ability of integrating different cues of information. When applying particle filter to a specific task, one key component need to be carefully defined is dynamic models, which characterize the motion of the target and determines how the particles are

propagated in the state space. Only when the particles are properly propagated, satisfied posterior may be obtained sequentially.

Previous works either choose complex dynamic models or simple dynamic models to characterize the target dynamics and to propagate particles [4,6,8,9]. The advantages of simple models are that they can be easily obtained and adapted to a specific application. Compared with complex models, simple models often show more elastic and robust with respect to noise since the states that can be reached are not carved tightly. Nevertheless, for high-dimensional tracking problems, simple models result in the most of particles with low weights and the efficiency of computation is low. As the dimension goes high, the exponential increasing computational burden quickly becomes prohibitively high to prevent simple models into practical use.

On the other hand, complex dynamic models incorporate more specific knowledge of how the object behaves than simple dynamic models. Therefore, it is more suitable for the high-dimensional tracking problems since computational cost is the key factor. Nevertheless, the complex dynamic models are typically learned from training examples or handcrafted using empirical knowledge. They are therefore very specific to the given task and are not easy to be obtained and adapted. Therefore, complex dynamic models are often learned by restricting the range of movement of the object and are easy to violate from the truth, e.g. assuming only walking or cycle motion can be handled for human body pose estimation [10]. These restrictions greatly reduce the generality of the resulting trackers.

In this paper, we propose a practical way to approaching the high-dimensional tracking problems which cumbered by complex motions. The main contributions of this paper can be concluded as follows:

First, we propose to using a set of simple motion models to approximating original strong motion models to ease the high dimensional curse.;

Then, we provide a practical solution of how to do inference by integrating multiple simple models in the particle filter.

Finally, we apply our proposed MMPF to the head motion tracking application. The experimental results show that the MMPF works well to this high-dimensional head motion tracking problem with reasonable computational resource.

The rest paper is organized as follows: In Section 2, We propose to approximate complex motion models using a set of simple motion models and to do integrated inference under the particle filter framework. We give the experiment results in Section 3 and conclude our work in Section 4.

2 Multi-model Particle Filter

Particle Filter is a technique for implementing a recursive temporal Bayesian filter by Monte Carlo simulations. The key idea is to represent the required posterior by a set of random samples and their associated weights. As the number of samples becomes sufficiently large, this Monte Carlo characterization becomes

an equivalent representation to the usual functional description of the posterior, and the particle filter approaches the optimal Bayesian estimate.

The power of the particle filter is in that it maintains a pool of hypotheses by sampling the proposal distribution $P(x_{i+1}|x_i)$ under the Bayesian framework. Generally, the more the hypotheses, the more chances to get accurate tracking results but the more computational resource is required. As the state of the object goes high, the computational cost quickly becomes prohibitively heavy due to the exponential computational complexity.

Following analysis can make this problem clear. To evaluate the efficiency of some particle set $\{x_i, \pi_i | i = 1, \dots, n\}$, two measurements are defined [4]. One is the survival diagnostic $D = (\sum_{i=1}^n \pi_i^2)^{-1}$, another is the survival rate $\alpha \approx \frac{D}{n}$, where n is the number of particles. To guarantee the performance, it can be inferred that the required number n of particles should be $n \geq \frac{D_{min}}{\alpha^d}$, where D_{min} is the minimum acceptable survival diagnostic considering performance. It's clear that α^d is the determining factor of required particle number n and where the computational difficulties mainly arise from the dimension d . Therefore, directly apply the particle filter to high-dimensional tracking problem is computational intractable.

The particle filter's property of generating a set of hypotheses provides a natural way to approximating complex motion model using a set of simple motion models to generate several kinds of hypotheses in the pool instead of only one kind: instead of propagate all particles using the original complex motion model, the particle set are branched and each sub set of particles are propagated using one simple motion model. The final result is obtained by composite all the estimates using graphical model probabilistically.

In the following paragraph, the proposed MMPF is described in detail mathematically. We first define the following terms:

P_{mn}^{i-} : The probability at time i that the complex motion will be explained from simple model m to simple model n due to variation of target dynamics. These probabilities are assumed to be known in prior here and satisfy $\sum_{m=1}^M P_{mn}^{i-} = 1$, where M is the number of simple models. A state transition matrix M_{mat}^{i-} , which stacks the P_{mn}^{i-} , combines M simple models according to a graphic model under the Markov assumption:

$$M_{mat}^{i-} = \begin{bmatrix} P_{11}^{i-} & \dots & P_{1M}^{i-} \\ \dots & \dots & \dots \\ P_{M1}^{i-} & \dots & P_{MM}^{i-} \end{bmatrix}. \quad (1)$$

P_{mn}^{i+} : The conditional probability that the target dynamics was explained from simple motion model m to simple motion model n at time i . Previous two probabilities describe how the simple models interact with each other to explain the complex motion model together.

P_m^{i-} : The probability that the target's dynamic will be explained by simple model m during time interval $[i, i + 1)$ and satisfy $\sum_{m=1}^M P_m^{i-} = 1$.

P_m^{i+} : The probability after simple models' interaction that the target dynamics can be explained by simple model m and satisfy $\sum_{m=1}^M P_m^{i+} = 1$.

Let $s_i = \{x_i^k, w_i^k, m_i^k | k = 1, \dots, N\}$ denote a particle set at time i , where m_i^k means the simple model according to which the particle k evolves in the state space at time i . For each particle, we define its private dynamic model according to the model probability P_m^{i-} and approximately there have the relation that the number of particles that will translate according to the simple model m is proportional to P_m^{i-} . That means all particles are divided into M groups probabilistically. Then, each group of particles behaves like a standard particle filter and M filtered states are obtained. Then the MMPF does an interaction between all filtered estimates and gets the final output by weighting all estimates statistically. After that, the model probability is updated according to the statistical property of residual error. The distance d_{i+1}^m which measures the residual error is application dependent and the distance we adopt to solve face tracking problem can be found in section 3.

Details of the algorithm are shown in Figure 1.

3 Application to Head Motion Tracking

In this section, we apply the MMPF method to head motion tracking with both rigid and non-rigid motions considered. Two difficulties are anticipated to be well handled under such a framework: one is that the method can work well with low quality image sequences and the other is that the tracker hold a high probability to recover from drift without manual re-initialization.

Experiments are performed on the real videos to test the tracker's ability of conquering clutters. The difficulties of tasks lie in that the states of the head need to be tracked are as high as 66 dimensions, which make the task is very challenging. While using MMPF, the original high-dimensional motions are factorized into eight simple models and make the problem tractable.

The experimental results show that the merits of this method can be concluded as follows:

- 1). The MMPF can be deal with low quality images due to the top-down matching scheme and stochastic search scheme (note that the experimental data we use are recorded using common hand held cameras and it was not high quality).

- 2). It is very robust to clutter. Even several frames are not well estimated and drift happen, the tracker holds a high probability to recover from the error. This is the essential merit for long sequence tracking in heavy clutter.

3.1 Face State Representation

A MPEG-4 compatible 3D parametric head model is implemented for synthesizing photo-realistic facial animations. One set of parameters can totally control the head motion and facial animations, named as Facial Animation Parameter (FAP) [11]. Here 66 low level FAPs are adopted instead of all 68 FAPs. In the experiments, we have made a try to use the CONDENSATION to do inference

on the 66 dimensional state space directly and stabilized results are not obtained even 10^8 particles are employed. The main reason is that most particles are wasted to generate useless hypotheses due to poor guidance. [2] has pointed out that when in spaces of dimensions much greater than about 10, good results are extremely difficult to get.

Iterate

Prediction: Sample N_m particles $s_i = \{x_i^k, w_i^k, m_i^k = m | k = 1, \dots, N_m\}$ from simple model $P_m(x_t|x_{t-1})$, satisfying that $\frac{N_m}{N}$ is proportional to P_m^- and $\sum_{m=1}^M N_m = N$.

Verification: Evaluate weights of N_m particles according to the likelihood model $w_t^k = P_m(y_t|x_t^{k,-})w_{t-1}^k$.

Interaction: 1). Compute an estimated state $\hat{x}_t^m = \sum_{j=1}^{N_m} \frac{w_j^k x_j^k}{w^k}$ for sub-model $P_m(x_t|x_{t-1})$. 2). Compute the model probability $P_m^{t+} = \sum_{m=1}^M P_{mn}^- P_m^{t-}$. 3). Compute the particle translation probability $P_{mn}^{t+} = P_{mn}^- P_m^{t-} / P_m^{t+}$. 4). Compute M filtered estimated states $\tilde{x}_{t+1}^n = \sum_{m=1}^M P_{mn}^{(t+)+} \hat{x}_{t+1}^m$. 5). Then the final result is $x_{t+1} = \sum_{n=1}^M P_n^{(t+)+} \tilde{x}_{t+1}^n$.

Updating Mode Probability: Compute distance d_t^m according to $P_m(z_t|x_t)$ for each model m and update its probability $P_m^{(t+1)-} = V_{t+1}^m P_m^{t+} / C$, where $V_{t+1}^m = \frac{\exp(-(d_{t+1}^m)^2/2)}{\sqrt{(2\pi)^R \sigma_{t+1}^m}}$ and C is a normalizing constant.

Re-sampling: Compute the covariance of the normalized weights. If this variance exceeds some threshold, then construct a new set of samples by drawing, with replacement, N samples from the old set, using the weights as the probability that a sample will be drawn. The weight of each sample is now $\frac{1}{N}$.

Fig. 1. The process of Multi-Model Particle Filter

Since previous experiment denied the particle filter with the original high-dimensional state representation as a practical solution. We first do a dimension reduction using PCA to get intrinsic representations of the face motion state with that the head pose parameters are canceled out by setting them to zero. The first five eigen-values in descending order are retained to accommodate 99% variation of the training data set. (The training data are obtained by manually turned). The head pose dynamics are modeled using three Nearly Constant Velocity Models (NCVM). Therefore, the final state is the coefficients of the

five eigen-vectors, which characterize facial expressions and three coefficients of NCVM models, span an eight dimensional sub-space. We combine two kinds of simple dynamic models to approximate the original complex motion models in the MMPF algorithm.

3.2 The Set of Simple Dynamic Models

In previous sub-section, five eigen-vectors obtained by dimension reduction technique and three NCVMs which corresponding to head yaw, tilt and roll respectively are chosen as sub models in the MMPF. Consequently, the original unknown motion model is factorized into eight simple models and each simple model varies only in one dimension,

$$P(x_i|x_{i-1}) = \sum_{m=1}^8 w_m P_m(x_i|x_{i-1}) \tag{2}$$

where $P_m(x_i|x_{i-1})$ represents one simple model and $P(x_i|x_{i-1})$ is the original strong motion model.

One step of the estimation process in experiments can be roughly represented by figure 2. The simple models are chained to do estimation in a cascade manner and construct a degenerate case of the MMPF. The previous estimated weight w_{m-1} of one simple model provides a starting point for the weight w_m 's estimation of next sub-model on the chain.

The model interaction matrix is assumed to be constant in the estimation process and set as that the diagonal elements of the matrix are 0.72 and the non-diagonal elements are set to 0.04 (In our experiments, the results are not sensitive to the small variation of these parameters). The prior probability P_m^- is initially set to $\frac{1}{M}$. In the tracking process, the distance measuring residual error of simple model m in frame $i + 1$ is set to

$$d_{i+1}^m = \frac{1}{|w_m^i - w_m^{i-1}|} \tag{3}$$

where w_m^i is the estimated weight of the simple model m in frame i .

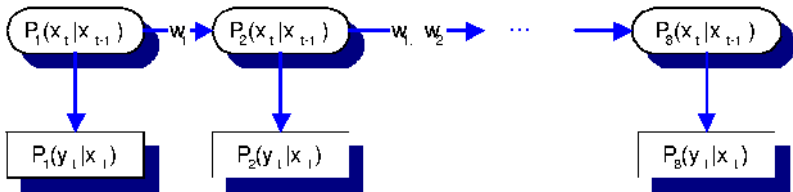


Fig. 2. The inference structure used for face tracking problem

3.3 Evaluating the Particles' Weights

In this sub-section, the likelihood model is constructed to relate the face states and face images and particles' weights are evaluated. When a frame I_t comes, the different image $\Delta I_t = I_t - I_{t-1}$ is first computed with the non-face area segmented out, where the I_{t-1} is the previous frame (all images are aligned manually according to the key feature points). Furthermore, ΔI_t is normalized like a matched filter to satisfy that: $\|\Delta I_t\| = 0$ and $var(\Delta I_t) = 1$. Then, the tracker propagates particles to generate a pool of hypotheses. For each new hypothesis h_t^i generated by particle k_i , an observation image O_t^i is generated by the previous mentioned 3D face model. Also, a difference image ΔO_t^i between the O_t^i and I_{t-1} is computed considering the face area for each O_t^i . Then a dot product between ΔO_t^i and ΔI_t is calculated as the corresponding particle's weight

$$w_t^i = \Delta I_t \cdot \Delta O_t^i = \langle \Delta I_t, \Delta O_t^i \rangle \quad (4)$$

3.4 Qualitative Performance Evaluation Using Real Video Data

Four video footages corresponding to four persons' facial animation are recorded to test our algorithm's ability to conquer clutter and the performance under real world conditions. For the limit of space, only one video is shown in figure.3. It has 189 frames and is at the resolution of 320X240. Note that there are some difference between the reconstructed 3D face model and the person himself due to the reconstruction error. The stabilized results are obtained by employing 6800 particles. The top row of figure 3 shows the sample frames of the recorded video and the bottom row shows the corresponding re-synthesizing frames by estimated parameters.

To test the tracker's ability to conquer clutter, we also disturb the tracker's estimates with the noise during tracking. Averagely, for each coefficients, 18% of the 3σ violation from the right value parameters value, the tracker can quickly recover from the drift within three frames with the probability of 90.5% during 500 times tests, where the σ is the standard variance learned from the training set.



Fig. 3. Comparing original frames with re-synthesized frames

4 Conclusions and Future Work

In this paper, we propose a novel method to deal with the tracking problem suffered from high-dimensional complex motions. Do inference in Bayesian filter frame-work, more information can be incorporated under this framework to promote the performance of the tracker.

Future works includes:

- 1) Composing low level cues, such as optical flow or other motion estimation techniques to guiding how to propagate particles and thus accelerate the running speed of the system to achieve near real-time performance;
- 2) Using 3D facial morphable model [12] to automatic initialization of the head motion tracking system.

Acknowledgements. This research is partially supported by National Hi-Tech Program of China (No.2001AA114190 and No. 2002AA118010), National Nature Science Foundation of China (No. 60332010), and ISVISION Technologies Co. Ltd.

References

1. M. Isard and A. Blake, CONDENSATION – conditional density propagation for visual tracking, In *Internal Journal of Computer Vision*, 29, 1, 5–28, 1998.
2. David A. Forsyth, Jean Ponce, *Computer Vision: A modern approach*, published by Prentice Hall, 2002.
3. D. Comaniciu, V. Ramesh and P. Meer, Real-time tracking of non-rigid objects using Mean Shift, In *IEEE Proceedings of CVPR*, Hilton Head Island, South Carolina, Vol. 2, 142-149, 2000.
4. J. MacCormick and M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracker, In *ECCV*, Vol.2, pp.3-19, 2000.
5. C. Rasmussen and G. Hager, Probabilistic Data Association Methods for Tracking Complex Visual Objects, *IEEE Transactions on PAMI*, Vol. 23, No. 6, June 2001.
6. J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *IEEE Proceedings of CVPR*, Hilton Head, V II pp. 126-133, 2000.
7. Kiam Choo and David J. Fleet. People tracking using hybrid monte carlo filtering. In *IJCV*, 2001.
8. H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, Vol.2, pages 702-718, 2000.
9. Vladimir Pavlovic, James M. Rehg, Tat-Jen Cham, and Kevin P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *ICCV*, 1999.
10. Rohr, K. Human movement analysis based on explicit motion models, In *Motion-Based Recognition*, kluwer Academic Publishers, Dordrecht Boston, 1997, ch.8, 171-198.
11. J.Ostermann, Animation of Synthetic Faces in MPEG-4, *Computer Animation*, pp.49-51, June 8-10, 1998.
12. V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces”, In *Proc. of SIGGRAPH99*, 1999.

Human Activity Recognition in Archaeological Sites by Hidden Markov Models

Marco Leo, Paolo Spagnolo, Tiziana D'Orazio, and Arcangelo Distanto

Institute of Intelligent Systems for Automation - C.N.R.
Via Amendola 122/D, 70126 Bari, ITALY
{leo,spagnolo,dorazio,distante}@ba.issia.cnr.it

Abstract. This work deals with the automatic recognition of human activities embedded in video sequences acquired in an archeological site. The recognition process is performed in two steps: first of all the body posture of segmented human blobs is estimated frame by frame and then, for each activity to be recognized, a temporal model of the detected postures is generated by Discrete Hidden Markov Models. The system has been tested on image sequences acquired in a real archaeological site meanwhile actors perform both legal and illegal actions. Four kinds of activities have been automatically classified with high percentage of correct decisions. Time performance tests are very encouraging for using the proposed method in real time applications.

1 Introduction

Automatic recognition of human activities is one of the most important and interesting open area in computer vision. Automatic visual surveillance, multi-modal interfaces and automatic indexing of multimedia data are some of the most common and relevant applications of this research field.

In this paper we focus on the automatic surveillance of archeological sites. Monitoring archeological sites is becoming a crucial problem in order to preserve buried and unburied property from thefts and vandalic actions. Nowadays archeological sites are monitored by using passive systems based on a set of large view cameras sending the acquired streams to an headquarter where one or more people, looking at the monitors, have to detect suspicious behaviours.

A large portion of open literature is devoted to human activity recognition in limited know spaces where the subjects dominate the image frame so that the individual body components (head, hands, etc.) can be reliably detected. Detailed reviews of these works can be found in [6,7]. Few works dealt, instead, with the problem of human activity recognition in large areas. CMU's Video Surveillance and Monitoring (VSAM) project [1], MIT AI Lab's Forest of Sensors Project [2] and VIGILANT project [5] are three of the most appreciate examples of recent research efforts in this field. In [2], the patterns (cars and humans) and their activities are learned by motion analysis. In [1], measurements based on a simple skeleton of the target are used to distinguish running people from walking

ones. In [5] velocity and width-to-height ratio of the patterns (car and human) are supplied as input to an HMM procedure.

Other considerable works in this area, like Pfinder [3] and W4 [4], try to classify humans and their activities by detecting features such as hands, feet and head, tracking and fitting them to an a prior human model.

The analysis of the related works reveals that these algorithms for large area monitoring can recognize very simple activities like vehicle and person entering and exiting form a parking area, people running or walking and so on. The automatic recognition of these simple actions could not be adequate to meet the requirements of the automatic surveillance of an archaeological site.

In this paper we propose a new approach for human activity recognition that works on binary patches extracted from the images containing human blobs. At first the horizontal and vertical histograms of human blobs are computed and supplied as input to an unsupervised clustering algorithm in order to detect the human posture in each frame. Then a statistical approach based on Discrete Hidden Markov Models is applied to temporal modelling the sequence of detected postures and to discriminate between legal and illegal activities. The last point that has been addressed in this work concerns the ability of the method to recognize in a long test sequence the beginning of the known activities. We have used a sliding window that has been overlapped to the test sequence to extract a fixed length observation sequence provided to the behavior classification step. The proposed approach has been validated using 165 long test sequences acquired in a real archeological site.

In the rest of the paper, first a description of the proposed activity recognition approach is explained and then the experimental results obtained on image sequences acquired in a real archaeological site meanwhile actors perform both legal and illegal actions are reported.

2 Human Activity Recognition

The human activity recognition system proposed in this paper works on the binary patches containing the human blob. For this reason, a preliminary people segmentation algorithm is required. Since the description of this algorithm is beyond the scope of this paper, we refer to the significant work proposed in the last years [9,12,13].

The behavior classification algorithm executes two steps: first of all the human body postures have to be estimated in each frame and then the temporal sequence of detected postures has to be modelled by discrete HMMs. In the pose estimation step horizontal and vertical histograms of the binary shapes are evaluated and supplied as input to an unsupervised clustering algorithm named BCLS (Basic Competitive Learning Scheme) [11]. In this work the proximity measure among two postures Im_1 and Im_2 is calculated as follows:

$$D(Im_1, Im_2) = d_1(X_1, X_2) + d_2(Y_1, Y_2) \quad (1)$$

where d_1 and d_2 are the Manhattan distances between the horizontal and vertical projections respectively. In particular a modified version of the Manhattan distance has been implemented; it was defined as:

$$d_2(Y1, Y2) = \min \left(\sum_{j=0}^{DimY-1} |Y1(j) - Y2(j+1)| \right) \quad (2)$$

$$d_1(X1, X2) = \min \left(\begin{array}{l} \sum_{j=0}^{DimX-1} |X1(j) - X2(DimX1 - j - i)| \\ \sum_{j=0}^{DimX-1} |X2(j) - X1(DimX1 - j - i)| \\ \sum_{j=0}^{DimX-1} |X1(j) - X2(j+i)| \\ \sum_{j=0}^{DimX-1} |X2(j) - X1(j+i)| \end{array} \right) \quad (3)$$

where the minimum is evaluated when i changes respectively in the interval $[0, DimY-1]$ and $[0, DimX-1]$.

In this new definition the vertical and horizontal histograms of an image are compared, by the proximity measure, with all the translated (and mirrored for the horizontal) versions (on the left and on the right) of the same histograms of another one. The minimum values are taken as the proximity measure.

In this way the proximity measure becomes invariant to the translation and mirroring of the binary target in the scene. Using the proposed proximity measure, the BCLS algorithm groups the available training images and then it classifies unknown new images on the base of their relative distances with respect to the built prototypes.

The recognition of human behavior is then performed by fully connected HMM in order to statistically analyze the temporal sequence of detected postures. In this step the number of different postures determines the number of the HMM codebook symbols (i.e the possible state values M) and each activity is associated to an HMM: this means that the number of HMM is always equal to the number of different activities of interest. Otherwise, the number of states N is fixed experimentally.

In the training phase the parameters of each HMM are updated in order to maximize the output probability of the training sequences. The training procedure based on the multiple observation sequence proposed in [10] has been used. This training solution has been adopted considering that different people perform the same activity in different ways. The algorithm proposed in [10] expresses the multiple observation probability as a combination of individual observation probabilities. In particular we have implemented a generalizing Baum's auxiliary function and we have built an associated objective function using Lagrange multiplier method. For each different activity an HMM model λ_i has been generated. In the test phase unknown sequences are provided as input to the HMMs. The probability to have the activity A given the observation sequence X of postures is computed by evaluating the forward backward probability. A decision criterion based both on maximum likelihood measure:

$$A^* = \operatorname{argmax} P(X|\lambda_i) \quad (4)$$

and a set of proper thresholds to manage unknown behavior has been introduced. Indeed each HMM has associated a threshold equal to the minimum probability value obtained during the training phase. The sequence X of posture observations is labeled as activity A if both its corresponding HMM gives the maximum likelihood measure among the whole set of HMMs and at the same time this probability value is greater than the relative HMM threshold. If this second condition is not satisfied the observation X cannot be associated to any of the known activities and is labeled as unknown.

The length of each observation sequence supplied to the HMMs is fixed in both training and testing phases and it has to be experimentally evaluated. In the training phase the observation sequences are segmented by hand whereas in the testing phase a sliding window (of the same length of training sequences) is used to cover the whole acquisition sequence.

3 Experimental Results

The proposed human activity recognition approach has been tested on real sequences acquired in an archaeological site. The images were acquired with a static TV camera Dalsa CA-D6. In order to consider only significant frames for the activity recognition process we have sampled the acquisition sequence tacking two frames per second. The software was implemented by using Visual C++ on a Pentium III 1 Ghz and 128 Mb of RAM. The archaeological site considered is a wide country area where some legal or illegal activities need to be discerned. In particular illegal activities are executed by people that first probe the subsoil using simple tools (such as sticks, tanks) and then they excavate to dig up some attracting objects. The people segmentation algorithms produces for each person in the scene a binary patch of 175x75 pixels. Starting from these patches, the BCLS algorithm detects three kinds of different postures: “standing”, “squatting” and “bent”. One example of each detected posture can be found in figure 1. Sequences composed by a temporal succession of these three postures are supplied as input to the HMMs in order to identify 4 kinds of activities:

1. Walking
2. Probing the subsoil by a stick
3. Damping the ground with a tank
4. Picking-up some objects from the ground.

The first activity, the simpler one, is legal; while the remaining ones are more complicated and illegal. The figure 2 shows some frames for each of the possible sequences of the different activities. In particular it can be note that the second and the third activities are very similar: they are composed by sequences of the same two postures, but with different temporal variations. The statistical modelling step is then composed by 4 HMMs. Each HMM is associated with a different kind of activity and it is trained with three different examples (performed by different people) of the associated activity. The training set, composed by $4 \times 3 = 12$ sequences is not changed during all the experiments described



Fig. 1. Three fundamental postures classified in the archaeological site.

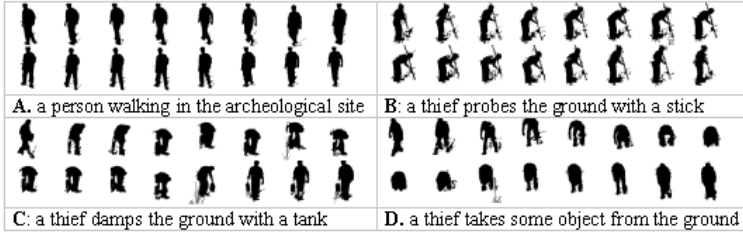


Fig. 2. Some frames extracted from 4 of the 12 sequences used to train the HMMs.

below. Each training sequence consists of 50 frames (so 50 is also the length of the sliding window used in the test phase). The experimental tests have shown that a greater number of training sequences decreases the generalization ability of the HMM, as asserted in [10].

In the first experiment, the system has been tested using 160 sequences. Each sequence contains one of the 4 activities to be recognized (just 40 for each kind of activity), but the beginning and the ending frames are not known. The length of the input sequence ranges from 400 to 1500 frames. If N_{TOT} is the total number of frame in each test sequence and n is the length of the sliding window then $N_W = N_{TOT} - (n - 1)$ is the number of windowed observation sequences O_w supplied as input to the HMMs for each test sequence.

An activity is recognized in a test sequence when at least one of its observation sequence O_w extracted by the sliding window, satisfies the recognition procedure described in the previous section (bayesian criterion + adaptive threshold).

In table 1 HMMs with 2-3-4-5-6-7-10 and 12 states have been tested in order to determine the optimal number N of HMM states in our application domain.

Each test sequence is given as input to the four HMMs with the same number of hidden states. For this reason, the results of every row of the table have to be considered altogether. The last column shows the mean percentage of correct classification; it sums up the best classification results obtained with 4 and 6 hidden states. In this case the percentage of right classification is 84.37%. HMM with a larger number of states have not been considered because the HMM's theory [8] suggests the use of a number of states much smaller than the number of symbols in each observation sequence (50 in our case). In the second experiment, in order to further improve the classification results, the four HMMs with the best classification performances have been selected and tested on the same 160

Table 1. The activity recognition results when the number of HMM states changes

HMM States	Activity									
	Walking People		Probing People		Damping People		Picking-up People		% of correct classification	
2	40/40	100%	40/40	100%	0/40	0.0%	40/40	100%	75	
3	40/40	100%	25/40	62.5%	29/40	72.5%	40/40	100%	83.75	
4	40/40	100%	25/40	62.5%	30/40	75%	40/40	100%	84.375	
5	40/40	100%	28/40	70%	26/40	65%	25/40	62.5%	74.375	
6	40/40	100%	27/40	67.5%	28/40	70%	40/40	100%	84.375	
7	40/40	100%	27/40	67.5%	26/40	65%	40/40	100%	83.125	
10	40/40	100%	20/40	50%	31/40	77.5 %	40/40	100%	81.875	

Table 2. The activity recognition results when the best hmm architecture of the exp.1 was used

Walking person HMM with 2 states		Probing people HMM with 2 states		Damping people HMM with 10 states		Pickig up people HMM with 2 states		Mean Percentage of corr. classification	
40/40	100%	29/40	72.5%	30/40	75%	40/40	100%	139/160	86.87%

sequences used in the experiment 1. Notice that in this case the performances of the proposed approach can change with respect to the ones reported in table 1, since both the relative maximum and corresponding threshold are used to classify each sequence. For the sequences “walking people” and “Picking up People” two hidden states have been selected because, under the same conditions, a smaller number of states makes simpler the training and test algorithms. For the sequence “Probing people” two states have also been selected because this case is the only one that ensures a classification performance of 100%. For the sequence “Damping People” the HMM with ten states has been selected since it ensures the best classification performance (77.50%).

The classification values relative to the selected HMMs are reported in cursive and bold type in table 1. The mean percentages of correct recognitions of the experiment 2 are reported in table 2 whereas table 3 shows the relative scatter matrix. The results demonstrate the effectiveness of the proposed approach based on a combination of HMM with different state numbers. The scatter matrix shows that the system mistakes the activities “probing people” and “Damping People”. Actually these two activities are very similar and hard to distinguish also for a human beings.

A further experiment was performed: we have supplied to the HMM architecture used in the experiment 2 a set of 5 sequences containing none of the 4 activities used in the training phase. In this case no false positives have been found (meaning that the threshold constraint relative to the winner HMM is never satisfied).

Finally, in order to evaluate the possibility of using the proposed approach for real time applications, some considerations about the computational load have

Table 3. Details of the activity recognition results when the best HMM architecture of the exp. 1 was used

Scatter Matrix	HMM Classification			
	Walking People	Probing People	Damping People	Picking up People
Walking Person	40	0	0	0
Probing Person	0	29	11	0
Damping Person	0	10	30	0
Picking up People	0	0	0	40

Table 4. Distribution of the computational load

Segmentation	Pose Estimation	Activity Recognition	Estimated Total Time per frame
$\sim 4 \times 10^{-2}$	$\sim 4 \times 10^{-2}$	$\sim 5 \times 10^{-5}$	$\sim 14 \times 10^{-2}$

been done. Each frame can be processed in about 14×10^{-2} s and the distribution of the computational load in the four subsystems is reported in table 4. The total amount allows the processing of 6 frames/sec. This can be a satisfying result taking in account that normally the human movements are slow.

4 Conclusions and Future Works

In this paper we have presented a reliable approach to recognize complex human activities performed by human beings in an archeological site. In particular we have addressed some of the problems concerning this kind of application domains.

Starting from the detection of moving people, the proposed approach addresses the problem of recognizing four different activities from temporal variations of postures. The postures have been detected using an unsupervised clustering algorithm that is able to separate the binary shapes in the required number of classes. Fixed length sequences (50 frames) of postures have been used both in training and test phase to model the four different activities and to classify new examples of the same behavior.

The experiments have demonstrated the effectiveness of using HMM to recognize activities based on sequence of temporal postures. Besides, the computational times have been evaluated for each step of the whole system: they are very encouraging for using the system in real time applications. Future work will be addressed to evaluate how a larger number of postures can improve the results of the activity classification, also considering that the same position of a person can be perceived in a different way from the camera according to the relative orientations. Besides, we will face the problem of selecting variable length observation sequences from the test sequences, in order to overcome the constraint imposed in this work of having the same behavior in quite the same number of frames.

References

1. R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, "A System for Video Surveillance and Monitoring", Technical Report CMU-RI-TR-00-12, Carnegie Mellon University, 2000.
2. C. Stauffer and W.E.L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking" IEEE transactions on PAMI, vol. 22, n.8, pp. 747-757, August 2000.
3. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pffinder: Real-time tracking of the human body. IEEE Transactions on PAMI, 19(7):780-785, 1997.
4. I. Haritaoglu, D. Harwood, and L. S. Davis, "W-4: Real-time surveillance of people and their activities," IEEE Transactions PAMI, vol. 22, no. 8, pp. 809-830, 2000.
5. P. Remagnino and G.A. Jones, "Classifying Surveillance Events from Attributes and Behaviors" in the Proceeding of the BMVC, Sept. 10-13, Manchester, pp. 685-694,2001.
6. D. Ayers and M. Shah, "Monitoring human behavior from video taken in an office environment", Image and Vision Computing, Vol. 19 (12) (2001) pp. 833-846.
7. M. Petkovic, W. Jonker and Z. Zivkovic, "Recognizing Strokes in Tennis Videos Using Hidden Markov Models", In proceedings of VIIP, Marbella, Spain, 2001.
8. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Processing", Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.
9. T. Kanade, T. Collins and A. Lipton, "Advances in cooperative multi sensor video surveillance, DARPA Image Understanding Workshop, Morgan Kaufmann, Nov. 1998, pp.3-24.
10. X. Li, M. Parizeau, R. Plamondon, "Training Hidden Markov Models with Multiple Observations - A Combinatory Method", IEEE Trans. on PAMI, vol. 22,4, pp.371-7, Apr.2000.
11. S. Theodoridis, K. Koutroumbas, "Pattern Recognition", Academic Press, San Diego, 1999, ISBN 0-12-686140-4.
12. A. Branca, G. Attolico, A. Distante "Cast Shadow Removing in Foreground Segmentation". In Proc. Int. Conf. on Pattern Recognition, 2002.
13. M. Leo, G. Attolico, A. Branca, A. Distante "People detection in dynamic images" In the proceedings of the IEEE WCCI, Honolulu, Hawaii, May 12-17, 2002.

An HMM Based Gesture Recognition for Perceptual User Interface*

HyeSun Park¹, EunYi Kim², SangSu Jang¹, and HangJoon Kim¹

¹ Department of Computer Engineering, Kyungpook National Univ., Korea
{hspark, ssjang, hjkim}@ailab.knu.ac.kr

² Department of Internet and Multimedia Engineering, Konkuk Univ., Korea
eykim@konkuk.ac.kr

Abstract. This paper proposes a novel hidden Markov model (HMM)-based gesture recognition method and applies it to the HCI to control a computer game. The novelty of the proposed method is two-folds. First one, the proposed method uses a continuous sequence of human motion as an input of HMM, instead of isolated data sequences or pre-segmented sequences of the data. The other one, it performs both segmentation and recognition of the human gesture automatically. To assess the validity of the proposed method, we applied the proposed system to a real game, Quake II, and then the results demonstrate that the proposed HMM can provide very useful information to enhance the discrimination between the different classes and reduce the computational cost.

Keywords: Gesture Recognition, HMM(hidden Markov model), User Interface.

1 Introduction

As a first step towards a perceptual user interface, many gesture recognition methods using computer vision techniques are developed. The gesture information is used in many application systems for alternative input tools instead of mouse, to provide a more convenient interface between user and computer [1, 2,3,4,5,6]. There are many existing techniques for gesture recognition using dynamic programming (DP) and HMMs. Among these, the HMM has attracted increasing attention for use in gesture recognition.

The HMMs have been widely used for many classification problems, as well as a gesture recognition problem. One of the main advantages of HMMs is their ability to model non-stationary signals or events. In the case of gesture, the signal is represented by the measurements of the body motion. This signal is non-stationary in nature: the display of a certain gesture in video is represented by a temporal sequence of body motions. Therefore it is natural to model each expression using an HMM trained for that particular type of expression.

* This research was supported by grant No.R05-2004-000-11494-0 from Korea Science & Engineering Foundation.

However the main problem with the HMMs in the gesture recognition is that it works on isolated data sequences or on pre-segmented sequences of the data from the video. In reality, this segmentation is difficult or unresolved, therefore there is need to find an automatic way of segmenting the sequences.

Some methods used a heuristic method on changes in the motion of several regions of the human body parts to decide that a gesture is beginning and ending. After detecting the boundaries, the sequence is classified to one of the gestures using the gesture-specific HMM. However, this method is prone to errors because of the sensitivity of the classifier.

Accordingly, to solve the above mentioned problems, we propose a new architecture of a HMM that can automatically segments and recognizes human gesture from video sequences. The proposed method performs both segmentation and recognition of the human gesture using a single HMM that composed of small HMMs for each gesture independently. There are thirteen HMMs, one for each gesture: Walk forward, back pedal, attack, turn left, turn right, step left, step right, look down, look up, center view, up/jump, down/crouch, and run. The proposed HMM takes a continuous stream of pose symbols as an input. Here, the pose is composed of three position vectors that indicate face, left hand and right hand, respectively. Each HMM continuously updates its state probabilities, whenever a new input pose arrives, and then recognizes a gesture when the probability of a *distinctive state* exceeds a predefined threshold value.

To assess the validity of the proposed HMM, it was applied to computer game. The experimentation shows that the proposed method can not only improve the accuracy of recognition but also reduce the computational cost when compared with a conventional HMM.

The rest of the paper is organized as follows. We describe the pose extraction and gesture recognition in Section 2 and Section 3, respectively. The experimental results and performance evaluation are presented in Section 4. Finally, Section 5 summarizes the paper.

2 A Pose Extraction

A certain gesture in real scene is represented by a temporal sequence of body motions. In this paper, we describe a gesture using motion of face and hands. Then, the position of these body parts at a specific time is called a *pose*. That is, a pose is to indicate the positions of face, left and right hands, thus it is represented as a vector $P = (F_x, F_y, L_x, L_y, R_x, R_y)$, where each element represents x -coordinate and y -coordinate of face, and left and right hands, respectively. The pose is extracted from each frame in the sequence by three steps: skin-color region extraction, connected-component labelling, and template matching. This extraction process is illustrated in Fig. 1.

An input frame is firstly divided into skin-color regions and non-skin-color regions using skin-color model that represented by 2-D Gaussian model [7].

Second, the results are filtered using connected-component labelling, and then positions of face, left hand, and right hand are obtained from 1st momentum

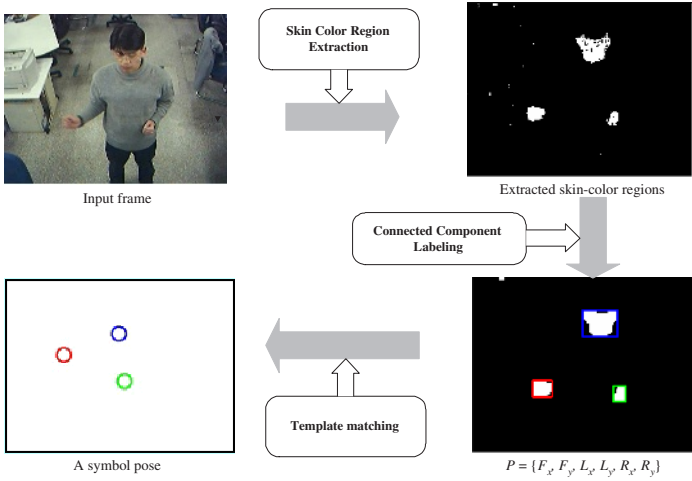


Fig. 1. The process of a pose extraction.

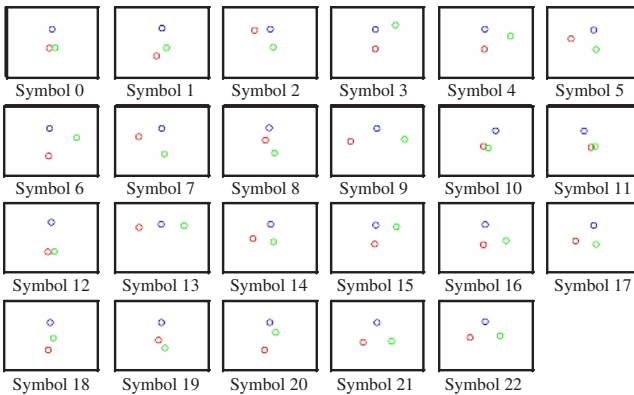


Fig. 2. A symbol table.

of the respective components. Finally, the extracted position vector is classified into the predefined pose symbol in a symbol table by a template matching: a position vector extracted from a frame is mapped to a pose symbol that has a smallest norm in the predefined symbol table.

Fig. 2 shows the pose symbols defined in this work. In this work, we assume that all gestures start and end with a same pose and that they can be distinguished by their distinctive pose. And the gesture has an intermediate pose between the start pose and its distinctive pose. Therefore a gesture is composed of four gestures: a starting pose, an intermediate pose, a distinctive pose, and ending pose. In Fig. 2, a symbol 0 is corresponding to a starting pose and ending

pose. And the symbols of from 1 to 13 are distinctive poses, and the others are intermediate poses.

3 Gesture Recognition

We use a HMM for gesture recognition. Traditional HMM usually works on isolated or pre-segmented sequences of input symbol sequences. Therefore there is a need to find an automatic way of segmenting the sequences. But in reality, it is hard to find an easy way to find this segmentation [8].

To solve the segmentation problem and enhance the discrimination between the classes, we propose a new architecture of HMMs. Fig. 3 shows the architecture of the proposed HMM.

The proposed HMM recognizes 13 gestures using a single HMM that composed of small HMMs for each gesture independently. The 13 gestures are that: Walk forward, back pedal, attack, turn left, turn right, step left, step right, look down, look up, center view, up/jump, down/crouch, run. These 13 gestures are basic command for interactive game. In this work, we assumed that all gestures start and end with a same pose and that they can be distinguished by their distinctive pose. Therefore, we can combine all gestures in a single model, as each small HMM shares a single ready state. In our model, each gesture state is composed of 4 states. Then, the proposed HMM recognizes a gesture from an input symbol stream using a 3rd state of each individual HMM as a path

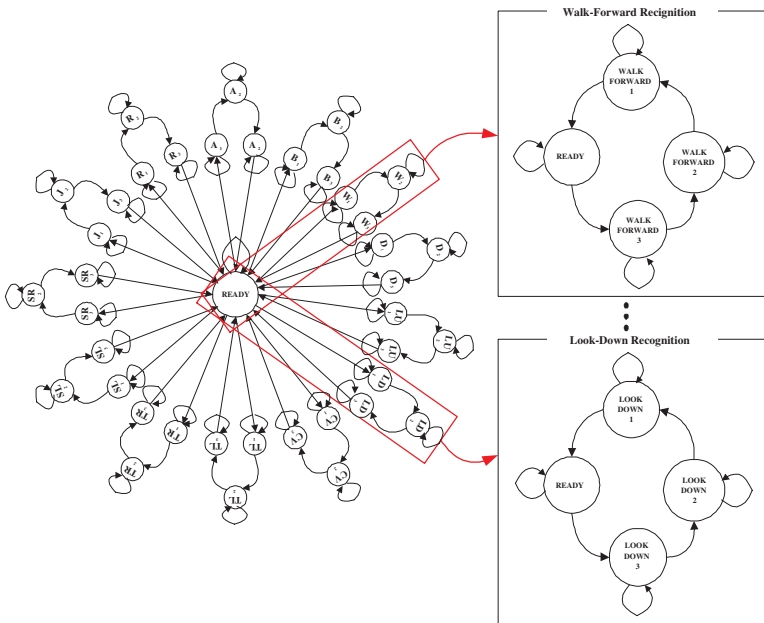


Fig. 3. An architecture of proposed HMM.

for the corresponding gesture. That is called *distinctive states*. And a gesture is detected and recognized, when the HMM passes this distinctive state.

The proposed HMM works as follows. First, the HMM starts with initial state probabilities $s^0 = s_k^0$, and continuously updates its state probabilities with the arrival of each input symbol as shown in Equation 1 and 2.

$$\underline{s}_n^t = \sum_{k=0}^{K-1} (s_k^{t-1} \times a_{kn}) \times b_{np} \quad (1)$$

$$s_n^t = \frac{\underline{s}_n^t}{\sum_{k=0}^{K-1} \underline{s}_k^t} \quad (2)$$

$S = \{s_k\}$: A vector of state probabilities, s_k denotes a state probability of state k .

K and M : The number of states and the number of poses, respectively.

$A = \{a_{ij}\}$: A $K \times K$ matrix for the state transition probability distributions, where a_{ij} is the probability of making a transition from state s_i to s_j .

$B = \{b_{ij}\}$: A $K \times M$ matrix for the observation symbol probability distributions, where b_{ij} is the probability of emitting pose symbol v_k in state s_i .

During this updating of its states, if a value of any distinctive state in one of each gesture has higher state probability than predefined threshold, a gesture includes that state is detected and recognized.

Consequently, the HMM described in this paper differs from the traditional HMM in two important respects, in its topology and behavior. The topological difference is that, for modelling the thirteen gestures, we use a single HMM rather than the more general approach of using 13 HMMs. Although the resulting HMM's topology is somewhat complex, it is not too complicated to design, as we can design a set of small HMMs for each gesture independently, and combine those into a single HMM. And for using a single HMM for recognition system makes it easier to utilize the relations between gestures, we hope this topology to allow us more systematic approach to the co-articulation problem. The behavioral difference is that, as an input of the HMM, we use a continuous symbol sequence, not the isolated candidate sequence selected by preprocessing stage. Although the state probability is updated for every input symbol, we expect less computing load for this updating requires much less computation compared to the traditional pre-segmented matching process. And the recognition system responds to the input in real time, not waiting for all the isolated candidate sequence selection.

In this paper, we assume that all of gestures share a ready state, namely, they model with relation of each gesture. Thus proposed HMM, which is the combination of the state sequence of the thirteen HMM's together, can provide very useful information and enhance the discrimination between the different classes.

4 Experimental Results

The proposed HMM based gesture recognition is implemented on Pentium IV using visual C++ language. Then, to show the effectiveness of the HMM based gesture recognition method, it is used as an interface of the computer game, Quake II.

For the experiments, we used a digital video camera located above the user at an angle of fifty degrees and a projector that projects the virtual world of the game onto a big screen. The test images are captured at a frame rate of 10 (Hz) and the size of each color image was a 320 × 240.

In Quake II, the frequently used commands are *walk forward*, *back pedal*, *attack*, *turn left*, *turn right*, *look up*, *look down*, *step left*, *step right*, *center view*, *up/jump*, *down/crouch* and *run*. These can be defined as gestures, which are shown in Fig. 4 and Fig. 5 shows an examples of pose sequences in some gestures.














Quake Command	Gesture	Quake Command	Gesture	Quake Command	Gesture
<u>Walk Forward (WF)</u> Shake a right hand in a forward direction.		<u>Back Pedal (BP)</u> Shake a left hand in a backward direction		<u>Attack (A)</u> Stretch out a right hand in front.	
<u>Turn Left (TL)</u> Stretch out a left hand to the left.		<u>Turn Right (TR)</u> Stretch out a right hand to the right.		<u>Look Up (LU)</u> Move a head to the left.	
<u>Step Left (SL)</u> Move a right hand down and a left hand to the left.		<u>Step Right (SR)</u> Move a left hand down and a right hand to the right		<u>Look Down (LD)</u> Move a head to the right.	
<u>Center View (CV)</u> Stretch out both hands horizontally, then return to the same position.		<u>Up/Jump (U/J)</u> Move both hands down and tilt head back.		<u>Down/Crouch (D/C)</u> Lift both hands simultaneously.	
<u>Run (R)</u> Swing arms alternately.					

Fig. 4. Thirteen types of gesture command used in Quake II.

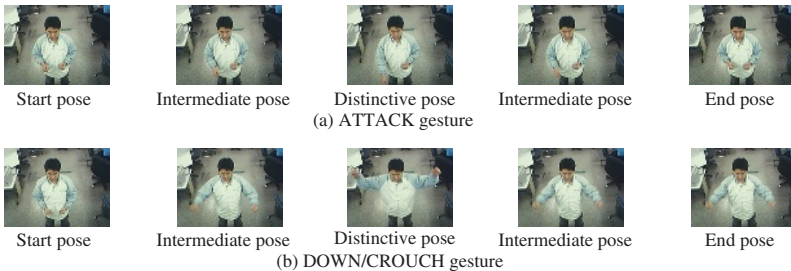


Fig. 5. Examples of pose sequences in some gestures.

Table 1. A gesture recognition result.

Input class	Results													
	A	BP	TL	TR	SL	SR	CV	DC	LD	LU	R	WF	UJ	Reject
A	157													3
BP	2	150												8
TL		2	154											4
TR				153										7
SL	2				156									2
SR						159								1
CV							159							1
DC								157						3
LD									158					2
LU		1								150				9
R											154			6
WF												155		5
UJ													159	1
Total error														52

Table 2. Performance comparison.

Gesture Types	Conventional HMM	Proposed HMM
Attack (A)	94.25%	98.13%
Walk Forward (WF)	88.75%	93.75%
Back Pedal (BP)	90.00%	96.25%
Turn Left (TL)	91.85%	95.63%
Turn Right (TR)	91.88%	97.50%
Run (R)	80.62%	99.38%
Step Left (SL)	96.12%	99.38%
Step Right (SR)	96.25%	98.13%
Look Up (LU)	86.75%	98.75%
Look Down (LD)	76.25%	93.75%
Center View (CV)	96.25%	96.25%
Up/Jump (UJ)	96.15%	96.88%
Down/Crouch (DC)	96.25%	99.38%

Table 3. Time comparison (10 Hz).

	Conventional HMM	Proposed HMM
Recognition time	3.19	1.84

Table 1 shows the performance of the proposed HMM in recognizing 13 commands. The result shows recognition rate of about 97.17% for the thirteen gestures. Table 2 shows the recognition rate of each gesture to compare a conventional HMM. The results indicate that the proposed HMM gave better results than the conventional HMM.

To assess the validity of the proposed HMM, the results were compared with those of conventional HMM, and then the results are summarized in Table 2 and Table 3. As you can see in Table 2, the proposed method can produce more accurate recognition rate. As well as, proposed HMM is averagely about 1.7 times faster than a conventional HMM.

Consequently, our experimentation shows that the proposed HMM have a great potential to a variety of multimedia application as well as computer games.

5 Conclusions

In this paper, we proposed a gesture recognition method to provide a more convenient user interface. For this, a new architecture of HMMs that automatically detects and recognizes human gestures from a continuous image sequence was developed. The proposed HMM differs from the most traditional HMMs because it works on a continuous input symbol stream, not on a finite symbol sequence pre-selected as a gesture candidate. The proposed method was applied to Quake II, and then the experimentation showed that our HMM produced better results than a traditional HMM.

References

1. R. Sharma, V.I. Pavlovic, and T.S. Huang, "Toward multimodal human-computer interface," *Proceeding of the IEEE*, vol. 86, pp. 853–869, 1998.
2. V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual interpretation of hand gestures for human-computer interaction: review," *Pattern Analysis and Machine Intelligence, IEEE Transaction on*, vol. 19, pp. 677–695, 1997.
3. H. Yu, Z. Yuanxin, X. Guangyou, Z. Hui, W. Zhen, and R. Haibin, "Video camera-based dynamic gesture recognition for HCI," *Signal Processing Proceedings – ICSP '98*, vol. 2, pp. 904–907, 1998.
4. E. Benoit, T. Allevard, T. Ukegawa, and H. Sawada, "Fuzzy sensor for gesture recognition based on motion and shape recognition of hand," *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems – VECIMS '03*, pp. 63–67, 2003.
5. K. Oka, Y. Sato, and H. Koike, "Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems," *Automatic Face and Gesture Recognition*, pp. 411–416, 2002.
6. H.S. Yoon, B.W. Min, J. Soh, Y.I. Bae, and H.S. Yang "Image Analysis and Processing," *Proceedings, Int. Conference on*. (1999) pp. 969–974, 1999.
7. J. Yang, A. Waibel, "A real-time face tracker, Applications of Computer Vision," *WACV*, vol. 15, no. 1, pp. 142–147, 1996.
8. I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, pp. 160–187, 2003.

Vision-Based Sign Language Recognition Using Sign-Wise Tied Mixture HMM

Liangguo Zhang^{1,3}, Gaolin Fang², Wen Gao^{1,2,3},
Xilin Chen², and Yiqiang Chen¹

¹ Institute of Computing Technology, Chinese Academy of Sciences,
P.O. Box 2704, Beijing 100080, China

² Department of Computer Science, Harbin Institute of Technology, China

³ Graduate School of the Chinese Academy of Sciences, China

{lgzhang, glfang, wgao, xlchen, yqchen}@jdl.ac.cn

Abstract. In this paper, a new sign-wise tied mixture HMM (SWTM-HMM) is proposed and applied in vision-based sign language recognition (SLR). In the SWTMHMM, the mixture densities of the same sign model are tied so that the states belonging to the same sign share a common local codebook, which leads to robust model parameters estimation and efficient computation of probability densities. For the sign feature extraction, an effective hierarchical feature description scheme with different scales of features to characterize sign language is presented. Experimental results based on 439 frequently used Chinese sign language (CSL) signs show that the proposed methods can work well for the medium vocabulary SLR in the unconstrained environment.

1 Introduction

Sign language, as a kind of most grammatically structured gesture, is one of the most natural and primary ways of exchanging information for most deaf people. The goal of SLR is to provide an efficient and accurate mechanism to transcribe sign language into text or speech, so that communication between deaf and hearing society can be more convenient. On the other hand, the research of SLR can serve as a good basis for the development of gestural human-computer interface. Therefore, SLR, as one of the important research areas of human-computer interaction (HCI), has attracted more and more interest in HCI society.

Until now, SLR can be mainly classified into two classes according to the devices used to collect gestures data, i.e., datagloves-based SLR [1,2] and vision-based SLR [3]-[7]. In the former case, users' freedom of movement is greatly limited and the datagloves should be carefully maintained. While in the latter one, people can interact with computer in a more natural way, which is just the research goal of natural HCI. Therefore, we focus on vision-based SLR. Some previous researches of vision-based SLR are as follows. Matsuo *et al.* [3] implemented a system to recognize 38 words of Japanese sign language (JSL) with a stereo camera for recording three-dimensional movements. Starner *et al.* [4] presented two video-based systems for real-time recognizing continuous American sign language (ASL) sentences on a 40-word lexicon, where the first system

observed the user from a desk mounted camera and 92% word accuracy was obtained, while the second one mounted the camera in a cap and 98% accuracy with restricted grammar was achieved. Vogler *et al.* [5] used computer vision methods to extract the three-dimensional feature parameters of a signer's arm motions and applied HMM to recognize continuous ASL sentences with an accuracy of 89.9% on a vocabulary of 53 signs. Grobel and Assan [6] used HMM to recognize the isolated signs with 91.3% accuracy on a 262-sign vocabulary. They extracted the features from video recordings of signers wearing colored gloves. HMM was also employed by Bauer and Hienz [7] to recognize continuous German sign language (GSL) with a single color video camera as input. An accuracy of 91.7% was achieved in recognition of sign language sentences with 97 signs.

As reviewed above, many researchers are putting their efforts on the research of vision-based SLR and some progress has been made, however, there are still many problems that need to be solved. For instance, how to make sound descriptions of sign features to discriminate the similar signs is a challenging problem, which is especially important for the medium or large vocabulary SLR task. How to provide a robust estimation of model parameters and achieve efficient computation of probability densities is also significant, especially when training data is relatively not sufficient.

In this paper, vision-based medium vocabulary CSL recognition using SWTMM-HMM is implemented to address above problems. To extract the feature information more precisely and provide the capability for the medium vocabulary SLR, 2D computer vision techniques are employed with the aid of a pair of simple colored cotton gloves in the unconstrained environment. For feature characterization, sign features are characterized in three hierarchical phases from different scales. When recognizing CSL signs, the proposed SWTMMHMM is used to achieve the efficient and effective training and recognition.

The rest of this paper is organized as follows. Section 2 describes some key ideas of SWTMMHMM. Section 3 presents the SWTMMHMM-based SLR system. Section 4 shows some experimental results and comparisons. The conclusions are given in the last section.

2 Sign-Wise Tied Mixture HMM

HMM [8] has been successfully employed to speech recognition, and applied by more and more SLR researchers in recent years. Formally, an HMM λ consisting of N states s_1, s_2, \dots, s_N can be defined by its parameters as $\lambda = (\pi, \mathbf{A}, \mathbf{B})$. Here, π stands for the vector of the initial probabilities π_i of the system starting in state s_i . The matrix \mathbf{A} represents the matrix of state-transition probabilities a_{ij} from state s_i to state s_j . The matrix \mathbf{B} can be in discrete or continuous form. In the former case, \mathbf{B} denotes observation probability distribution of state s_i as $b_i(v)$, v is any discrete observation symbol and discrete HMM (DHMM) is determined. In the latter case, \mathbf{B} represents observation probability density function of state s_i as $b_i(\mathbf{X})$, \mathbf{X} is any continuous observation vector; Accordingly, continuous HMM (CHMM) is drawn. In fact, both DHMM and CHMM can be viewed as special forms of semi-continuous HMM (SCHMM) [9]

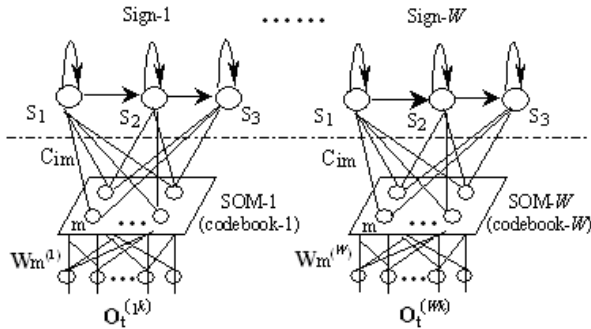


Fig. 1. Architecture of the SWTMHMM

or tied mixture HMM (TMHMM) [10]. In SCHMM, however, a global codebook is shared by all the HMM states of all words, which makes the amount of mixture weights too large to be robustly estimated, especially when the amount of training data is small to moderate.

Inspired by the work in speech recognition [11,12], we propose the SWTMHMM for SLR, where the mixture densities of the same sign model are tied so that the states belonging to the same sign share a common local codebook of Gaussian densities, other than a global codebook in SCHMM. Since Gaussian mixtures of each sign model are tied, SWTMHMM can obtain a tradeoff between a large number of mean vectors and covariances (e.g., in large-scale CHMMs) and an excessive amount of mixture weights (e.g., in large-scale SCHMM), and also SWTMHMM can be trained in a straightforward way.

2.1 SWTMHMM Architecture

Fig. 1 shows the architecture of the SWTMHMM, which consists of two functional layers, namely self-organizing map (SOM) [13] based VQ codebook generation layer and HMM modeling layer. Firstly, SOM is employed to generate the codebook of Gaussian densities for each initial sign model from the corresponding training samples. Then, each formed “local” codebook for each sign is coupled into the sign-wise tied mixture model training in the HMM framework.

Let the number of signs be W and the observation vector sequence of the sign w be $O_t^{(w)} = \{o_{td}, d = 1, 2, \dots, D\}$, where t is the frame number of the sequence and D is the dimension of the observation vector. As seen from Fig. 1, each O_t is connected to each neuron of each SOM with the weight vector $W_m^{(w)} = \{w_{md}, d = 1, 2, \dots, D\}$, i.e., after SOM training, we can regard $W_m^{(w)}$ as the code word $V_m^{(w)}$ and the generated “local” codebook will be shared by all states of the same sign model. Each neuron is linked to each state s_i of each sign model with the weight coefficient c_{im} . In this case, the observation probability of SWTMHMM in state s_i can be defined as

$$b_i(\mathbf{O}_t^{(w)}) = \sum_{m=1}^M c_{im} b_{im}(\mathbf{O}_t^{(w)}) = \sum_{m=1}^M c_{im} f(\mathbf{O}_t^{(w)} | \mathbf{V}_m^{(w)}), \quad (1)$$

where $f(\mathbf{O}_t^{(w)} | \mathbf{V}_m^{(w)})$ can be a Gaussian distribution and $\sum_{m=1}^M c_{im} = 1$.

2.2 SWTMHMM Parameter Estimation

Let K training samples for each sign be $\mathbf{O} = [\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, \dots, \mathbf{O}^{(K)}]$, where $\mathbf{O}^{(k)} = \{\mathbf{O}_t^{(k)}, t = 1, 2, \dots, T_k\}$ is the k th observation sequence, and T_k is the frame number of the sequence. The original model is denoted as λ and the re-estimated model $\bar{\lambda}$. Let the number of neurons of each SOM for each sign be M , we can get the neurons set (i.e., mean vectors of Gaussian codebook) $\mathbf{V} = [\mathbf{V}_m, m = 1, 2, \dots, M]$ for each sign after its corresponding SOM training. We denote \mathbf{S} as a state sequence $\mathbf{S} = s_1, s_2, \dots, s_{T_k}$, $s_i \in \{1, 2, \dots, N\}$, where N is the number of states for each sign model. We assume each observation sequence is independent of every other one, and the model training aims to maximize $P(\mathbf{O} | \lambda) = \prod_{k=1}^K P(\mathbf{O}^{(k)} | \lambda)$ by adjusting the parameters of λ . Thus, we may employ expectation-maximization (EM) algorithm to realize the maximum likelihood estimation (MLE) based optimization of λ by introducing the auxiliary function $Q(\lambda, \bar{\lambda})$, and iterating the following expectation step (E-Step) and maximization step (M-Step) for several passes until convergence:

E-Step: $Q(\lambda, \bar{\lambda}) = E[\log \prod_{k=1}^K P(\mathbf{O}^{(k)} | \bar{\lambda}) | \mathbf{O}, \lambda]$, *M-Step:* $\bar{\lambda} \leftarrow \arg \max_{\hat{\lambda}} Q(\lambda, \hat{\lambda})$.

$Q(\lambda, \bar{\lambda})$ can be defined as

$$Q(\lambda, \bar{\lambda}) = \sum_{k=1}^K \sum_{\mathbf{S}} \sum_{\mathbf{V}} \frac{P(\mathbf{O}^{(k)}, \mathbf{S}, \mathbf{V} | \lambda)}{P(\mathbf{O}^{(k)} | \lambda)} \log P(\mathbf{O}^{(k)}, \mathbf{S}, \mathbf{V} | \bar{\lambda}), \quad (2)$$

where

$$\begin{aligned} \log P(\mathbf{O}^{(k)}, \mathbf{S}, \mathbf{V} | \bar{\lambda}) &= \log \left(\prod_{t=1}^{T_k} \bar{a}_{s_{t-1}s_t} \bar{b}_{s_t v_t}(\mathbf{O}_t^{(k)}) \bar{c}_{s_t v_t} \right) \\ &= \log \bar{\pi}_{s_1} + \sum_{t=1}^{T_k-1} \log \bar{a}_{s_t s_{t+1}} + \sum_{t=1}^{T_k} \log \bar{b}_{s_t v_t}(\mathbf{O}_t^{(k)}) + \sum_{t=1}^{T_k} \log \bar{c}_{s_t v_t} \end{aligned}$$

Given λ and $\mathbf{O}^{(k)}$, we define the forward and backward variable as follows:

$$\begin{aligned} \alpha_t^{(k)}(i) &= P(\mathbf{O}_1^{(k)} \mathbf{O}_2^{(k)} \dots \mathbf{O}_t^{(k)}, s_t = i | \lambda) \\ \beta_t^{(k)}(i) &= P(\mathbf{O}_{t+1}^{(k)} \mathbf{O}_{t+2}^{(k)} \dots \mathbf{O}_{T_k}^{(k)} | s_t = i, \lambda) \end{aligned}$$

which can be computed by *forward-backward procedure* [8]. We further define the intermediate probability of being in state i at frame t with $\mathbf{O}_t^{(k)}$ quantized to \mathbf{V}_m as follows:

$$\begin{aligned} \xi_t^{(k)}(i, m) &= f(s_t = i, \mathbf{O}_t^{(k)} \rightarrow \mathbf{V}_m | \mathbf{O}^{(k)}, \lambda) \\ &= \frac{\alpha_t^{(k)}(i) \beta_t^{(k)}(i)}{\sum_{i=1}^N \alpha_t^{(k)}(i) \beta_t^{(k)}(i)} \cdot \frac{c_{im} b_{im}(\mathbf{O}_t^{(k)})}{b_i(\mathbf{O}_t^{(k)})} \end{aligned} \quad (3)$$

Therefore, we can maximize every item in $Q(\lambda, \bar{\lambda})$ by setting $\partial Q(\lambda, \bar{\lambda})/\partial \bar{c}_{im} = 0$ and $\nabla_{\bar{\mathbf{V}}_m} Q(\lambda, \bar{\lambda}) = 0$, and get the re-estimation formulas for \bar{c}_{im} and $\bar{\mathbf{V}}_m$ ($1 \leq m \leq M$) as follows:

$$\bar{c}_{im} = \frac{\sum_{k=1}^K \sum_{t=1}^{T_k} \xi_t^{(k)}(i, m)}{\sum_{k=1}^K \sum_{t=1}^{T_k} \sum_{m=1}^M \xi_t^{(k)}(i, m)} \quad (4)$$

$$\bar{\mu}_m = \bar{\mathbf{V}}_m = \frac{\sum_{k=1}^K \sum_{t=1}^{T_k} \sum_{i=1}^N \xi_t^{(k)}(i, m) \mathbf{O}_t^{(k)}}{\sum_{k=1}^K \sum_{t=1}^{T_k} \sum_{i=1}^N \xi_t^{(k)}(i, m)} \quad (5)$$

Similarly, $\bar{\Sigma}_m$ can be obtained as

$$\bar{\Sigma}_m = \frac{\sum_{k=1}^K \sum_{t=1}^{T_k} \sum_{i=1}^N \xi_t^{(k)}(i, m) (\mathbf{O}_t^{(k)} - \bar{\mu}_m)(\mathbf{O}_t^{(k)} - \bar{\mu}_m)'}{\sum_{k=1}^K \sum_{t=1}^{T_k} \sum_{i=1}^N \xi_t^{(k)}(i, m)} \quad (6)$$

The re-estimation formulas for π_i , a_{ij} are the same to the CHMM [8]. When training, the constraint should be imposed on c_{im} , namely $\sum_{m=1}^M c_{im} = 1$.

3 SWTMHMM-Based CSL Recognition System

The structure of the SWTMHMM-based CSL recognition system is shown in Fig. 2. which mainly consists of two modules, i.e., vision-based feature extraction and SWTMHMM-based recognizer train/recognition module.

3.1 Feature Extraction

In order to realize the vision-based medium vocabulary SLR, we propose a robust feature detection framework in the unconstrained environment and an effective hierarchical feature characterization scheme [16].

To describe the hand features more elaborately to discriminate the similar signs, a couple of colored cotton gloves are used, which doesn't limit the user's freedom of movement. Considering the fact that most of discriminative feature information is conveyed by the dominant hand (D-hand, usually the right hand

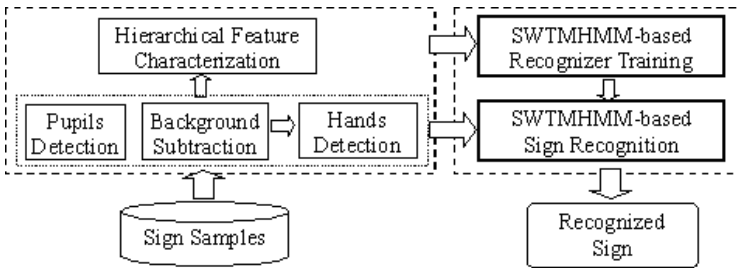


Fig. 2. Overview of SWTMHMM-based SLR system

for most of the people), we color the D-hand glove with 7 different colors to indicate the areas of 5 fingers, palm and back, while the non-dominant hand (ND-hand, usually the left hand) glove with another plain color.

For the vision-based detection, a pupils-detection algorithm [14] is applied to detect the positions of two pupils, which aims to provide a reference point to position. Meanwhile, an improved background subtraction algorithm [15] is used to remove the background from the frame image, which helps the system to work well in the unconstrained environment. After the body area has been roughly segmented from the background, a double-hands detection algorithm making use of both hand color information and shape geometry constraint information is designed to robustly detect both hands and an ellipse is applied to fit the ND-hand area. Two examples of vision-based detection results can refer to [16]. More examples of video detection results are available at <http://www.jd1.ac.cn/user/lgzhang/Research/VCSLR.htm>.

After the vision-based feature detection, we describe the sign features in the form of feature vectors in three hierarchical phases [16]. In the first phase, we mainly characterize the feature of the D-hand from shape, orientation and location in the form of a 35-dimension feature. In the second phase, we describe the ND-hand shape and orientation feature by using parameters of the ellipse fitting the area of the hand. And in the third phase, we employ PCA to further characterize the pixels distribution features of 5 fingers of the D-hand.

3.2 SWTMHMM-Based Training and Recognition

Based on the analysis of the SWTMHMM in the previous section 2, SWTMHMM-based training and recognition can be briefly described as follows.

SWTMHMM-Based Recognizer Training. The SWTMHMM can be trained like following.

- 1) For each sign, train the SOM from the corresponding training samples to generate the local codebook \mathbf{V} that initializes model parameters of mean vectors.
- 2) Initialize the parameters of π_i , a_{ij} , c_{im} , Σ_{im} .
- 3) Re-estimate the parameters with the re-estimation formulas and all observation sequences for the corresponding sign.
- 4) Terminate the procedure if the convergence criterion is met; otherwise, replace old parameters with the new ones, and return to 3).

SWTMHMM-Based Recognition. The problem of sign recognition is to choose a model that can best characterize the observation signal. Thus, the recognition procedure is as follows.

- 1) For the observation sequence $\mathbf{O} = \mathbf{O}_1\mathbf{O}_2 \cdots \mathbf{O}_T$, compute $b_i(\mathbf{O}_t)$ among all signs.
- 2) Decode with *Viterbi* algorithm [8] in term of $b_i(\mathbf{O}_t)$.
- 3) The result is the sign that has the maximum probability of decoding among all signs, i.e., $w_{result} = \arg \max_{1 \leq w \leq W} P(\mathbf{O}|\lambda_w)$.

4 Experiments

The experiment is performed in our laboratory with unconstrained background and the fluorescent illumination. Only a color camera is employed and placed in front of the signer to capture the sign video data, where the video is analyzed at 320 by 240 pixel resolution. The computer is a PC with an 800MHz Pentium III CPU and 256M RAM. Since the current system is signer-dependent, the training and test data are both captured by the same signer. The recognition vocabulary consists of 439 frequently used CSL signs for deaf people during the daily communication. Four samples of each sign are collected for training and one sample for test. Thus, totally 1756 training and 439 test samples are collected.

In order to test the performance of the proposed hierarchical feature characterization scheme, we first experiment the sign recognition by using CHMM. And the dynamic programming technique is employed to estimate different number of states for each HMM model; Thus, the average number of states for each sign is 4.87, and the number of mixtures for each state of each sign is 5. Then, the current CSL sign recognition system achieves an average recognition accuracy of 92.5% on 439 signs and the hierarchical feature description scheme effectively improves the recognition accuracy in a progressive manner [16].

To evaluate the performance of the proposed SWTMHMM, we compare that with CHMM and SCHMM. Table 1 gives the results of performance comparison. Here, each type of HMM is 3-states left-to-right sign model allowing possible skips. And the recognition time per word in the table do not include preprocessing, e.g., feature extraction, which is the same for all types of HMM.

From Table 1, we can see that the SWTMHMM provides more satisfying configurations when comparing the number of parameters, recognition speed and accuracy with that of CHMM and SCHMM. The reason for the above results is that, through sharing mixtures within each sign model, the number of mixture components and mixture weights is greatly reduced, which helps SWTMHMM to achieve robust parameters estimation and efficient computation of the probability densities.

5 Conclusions

This paper presents the SWTMHMM for the vision-based SLR. SWTMHMM can provide more satisfying configurations than CHMM and SCHMM by sharing the same codebook within the same sign model, which leads to robust parameters

Table 1. Performance comparison of different types of HMM

Type of HMM	Number of mixtures for each state (M)	Number of parameters		Recognition performance	
		weights(10^3)	means(10^3)	speed(s/word)	accuracy(%)
CHMM	5	6.6	6.6	0.165	91.3
SCHMM	1024	1,348	1.0	0.131	89.7
SWTMHMM	6	7.9	2.6	0.094	90.9

estimation and efficient computation of probability densities. In addition, by applying the techniques of hands detection, background subtraction and pupils-detection to detect the feature areas precisely with the aid of simple colored gloves, and employing an effective hierarchical feature characterization scheme, we implement a medium vocabulary CSL recognition system that can be applied in the unconstrained environment. Experimental results show that our methods work well for the medium vocabulary SLR task.

Acknowledgement. This work was supported by Natural Science Foundation of China under Grant 60303018.

References

1. Liang, R.H., Ouhyoung, M.: A Real-Time Continuous Gesture Recognition System for Sign Language. In: AFGR, (1998) 558-565
2. Gao, W. *et al.*: Handtalker: A Multimodal Dialog System Using Sign Language and 3-D Virtual Human. In: ICMI, (2000) 564-571
3. Mastuo, H. *et al.*: The Recognition Algorithm with Non-contact for Japanese Sign Language Using Morphological Analysis. Proc. GW, (1997) 273-284
4. Starner, T. *et al.*: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. IEEE PAMI, vol. 20, no. 12, (1998) 1371-1375
5. Vogler, C., Metaxas, D.: Adapting Hidden Markov Models for ASL Recognition by Using Three-Dimensional Computer Vision Methods. Proc. SMC, (1997) 156-161
6. Grobel, K., Assan, M.: Isolated Sign Language Recognition Using Hidden Markov Models. Proc. SMC, (1996) 162-167
7. Bauer, B., Hienz, H.: Relevant Features for Video-Based Continuous Sign Language Recognition. In: AFGR, (2000) 440-445
8. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, vol. 77, no. 2, (1989) 257-285
9. Bellegarda, J.R., Nahamoo, D.: Tied Mixture Continuous Parameter Modeling for Speech Recognition. IEEE Trans. ASSP, vol.38 (12), (1990) 2033-2045
10. Huang, X.D.: Phoneme Classification Using Semi-continuous Hidden Markov Models. IEEE Trans. Signal Processing, vol. 40, (1992) 1062-1067
11. Digalakis, V., Murveit, H.: Genones: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognition. In: ICASSP, (1994) 537-540
12. Kurimo, M.: Hybrid Training Method for Tied Mixture Density Hidden Markov Models Using Learning Vector Quantization and Viterbi Estimation. Proc. IEEE Workshop on NNSP, (1994) 362-371
13. Kohonen, T.: The Self-Organizing Maps. Proceedings of the IEEE, vol. 78, no. 9, (1990) 1464-80
14. Miao, J., Gao, W. *et al.*: Gravity-center Template Based Human Face Feature Detection. In: ICMI, (2000) 207-214
15. Jabri, S. *et al.*: Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information. In: ICPR, vol. 4, (2000) 627-630
16. Zhang, L.-G. *et al.*: A Vision-Based Sign Language Recognition System Using Tied-Mixture Density HMM. In: ICMI (2004), State College, Pennsylvania, USA

Author Index

- Abdel-Baki, Nashwa II-447
Abdelli, A. III-273
Abidi, Besma III-69
Abidi, Mongi III-69
Agnihotri, Lalitha II-9
Agusa, Kiyoshi II-188
Ahn, Byungha III-482
Ahn, Jinhaeng III-386
Ahn, Sang Chul II-665
Aizawa, Kiyoharu I-174, II-915
Akafuji, Tomohisa I-478
Ambikairajh, Eliathamby II-882
Amin, Tahir I-642
Ampornaramveth, V. I-369
Anyoji, Shigemitsu II-841
Ariki, Yasuo III-466
Arita, Daisaku I-418
Asano, Shoichiro II-770
Ashourian, Mohsen III-747
Au, Oscar C. III-314
- Babaguchi, Noboru I-626, II-121, II-272
Bae, Tae Meon I-335
Baek, Jang-Woon II-712
Baek, Yun-Cheol II-415
Bao, Hongqiang I-312
Barbieri, Mauro II-9
Barrera, Salvador III-1
Bashar, M.K. II-188
Bhagavathy, Sitaram II-256
Bhuiyan, M.A. I-369
Boujemaa, Nozha II-497
Bui, Son Hai II-264
Byun, Hyeran II-138
Byun, Youngbae III-331, III-721
- Cai, Jianfei III-167, III-566, III-673
Cai, Rui I-49, II-890
Cha, Jae-Ryong II-696
Cha, Jae-sang I-548
Cha, Jongeun III-482
Chan, Syin I-65
Chang, Chin-Chen III-731
Chang, Chun-Fa III-474
Chang, Feng-Cheng I-157, III-356
Chang, Fu-Mau III-151
Chang, Hung-Yi III-647
Chang, Hyeyoung II-364
Chang, Jing-Ying II-130
Chang, Juno II-415, II-601
Chang, Long-Wen II-50, III-410
Chang, Ray-I I-114
Chang, Yongseok I-165
Che, Lichang II-389
Chen, Chien-Hsung III-410
Chen, Duan-Yu II-221
Chen, Enyi II-754
Chen, Guanrong III-418
Chen, Homer I-239
Chen, Jie II-585
Chen, Jinji I-532
Chen, Liang-Gee II-130
Chen, Rong-Chang III-647
Chen, Xilin II-1035
Chen, Y.H. III-515
Chen, Yiqiang II-1035
Chen, Yung-Sheng I-327
Cheng, Qian-Sheng III-95
Cheng, Ren-Hao I-33
Cheng, Wen-Huang III-558
Chia, Liang-Tien I-65, I-182, I-470, I-634, II-993, III-566
Chiang, Tihao II-521, III-215
Chiu, Patrick I-73
Cho, Choong-Ho I-25
Cho, Jeong-Hyun I-165
Cho, Nam Ik III-257
Cho, Sung-Joon II-720
Cho, Wan-Hyun III-378
Cho, Yong-Bum II-720
Cho, Yong Ho III-630
Cho, Yongjoo II-601
Cho, Yookun I-572
Cho, You-Ze II-322
Cho, Youngran III-491
Choe, Wonhee II-786
Choi, Byeong-Doo II-529
Choi, Chang-Gyu I-165
Choi, Chang Yeol III-290
Choi, Ho-Seung II-696

- Choi, Inchoon II-967, III-386
 Choi, Jae-Gark III-615
 Choi, Jonguk III-331, III-508, III-721
 Choi, Kwang-Pyo II-967
 Choi, Seung Ho II-381
 Choi, Soo-Mi I-319, II-569
 Choi, Woong Il III-224
 Choi, Yanglim II-786
 Choi, Young-Soo II-322
 Chou, Chun-Hsien II-833
 Chu, Wei-Ta III-558
 Chuang, Wei-Hong I-239
 Chun, Kang Wook III-372
 Chung, Hawik III-207
 Chung, Kidong I-610
 Chung, Sang-Hwa I-618
 Crucianu, Michel II-497
 Cui, Chenyang III-103
 Cui, Ting II-338
- D'Orazio, Tiziana II-1019, III-655
 Dachuri, Naveen I-426
 Dai, Li-Rong II-431
 Dao, Minh-Son III-550
 Daoudi, M. III-273
 Day, Miin-Luen III-394
 da Silva Cruz, Luis A. II-545
 DeNatale, Francesco G.B. III-550
 Deng, Lei III-200, III-639
 Dimitrova, Nevenka II-9
 Ding, Wei III-426
 Ding, Xiaoqing I-57
 Distante, Arcangelo II-1019
 Divakaran, Ajay II-27
 Dou, Wenhua I-602
 Droese, Michael III-249
 Duan, Lijuan II-172
 Duan, Ling-Yu II-479, III-566
- El-Saban, Motaz II-256
 Eom, Jong-hoon II-356
 Eom, Seongeun III-482
- Fang, Gaolin II-1035
 Fang, Hung-Chi II-130
 Ferecatu, Marin II-497
 Foote, Jonathan I-73
 Fu, Hsin-Chia III-515
 Fujihata, Masaki III-8
 Fujii, Toshiaki I-497, III-249
- Fujikawa, Kazutoshi III-45
 Funatsu, Satoshi III-15
 Funayama, Ryuji II-736
 Furuya, Masanori I-271
- Gao, Bin III-95
 Gao, Wen II-172, II-513, II-585, II-985,
 II-1011, II-1035, III-192, III-200, III-433
 Gevers, T. II-229
 Gouet, Valerie III-61
 Großmann, Hans Peter II-447
 Gu, Jing II-890
 Gu, Liexian I-57
 Gu, Xiao-Dong II-431
 Guan, Ling I-131, I-642, II-898
 Guaragnella, Cataldo III-655
 Guo, Hui II-439
 Guo, Xun III-665
- Ha, Jong-Eun I-214, I-402
 Hagita, Norihiro III-87
 Hahn, Hernsoo III-207
 Hamada, Reiko II-657
 Hamamoto, Takayuki III-53
 Han, Hee-Seop II-688
 Han, Hui III-441, III-771
 Han, Sun-Gwan II-688
 Han, Youngjoon III-207
 Hang, Hsueh-Ming I-157, II-521, II-672,
 III-356, III-763
 Hanson, Allen II-778
 Harumoto, Kaname I-478
 Hasanuzzaman, Md. I-369
 Hasegawa, Shoichi III-31
 Hashimoto, Masayuki II-593
 Hashimoto, Naoki III-23
 Hasida, Kōiti I-478
 Hatori, Mitsutoshi II-728
 He, Jingrui I-198, II-213, III-111
 He, Li-wei I-105
 He, Sheng-Fang III-558
 Heath, Christian I-81
 Ho, Jan-Ming I-114
 Ho, Yo-Sung I-230, I-410, II-866, III-135,
 III-159, III-340, III-622, III-747
 Hong, Hyeonok I-610
 Hong, Kwangjin II-81
 Hong, Min-Cheol II-537, III-705
 Hong, Sang Hyuk II-113
 Hong, Suk-ki II-180

- Horiuchi, Takahiko I-190, II-794
 Horng, Mong-Fong I-33
 Hotta, Seiji II-162
 Hsiao, Dun-Yu I-239
 Hsu, Chih-Lung II-849
 Hsu, Wen-Hsing I-327
 Hu, Ming-Zeng III-200, III-639
 Hu, Shuo-Cheng III-647
 Hu, Yiqun I-470, II-993
 Hua, Xian-Sheng I-57, II-289, II-1001
 Hua, Zhigang II-704
 Huang, Feng II-947
 Huang, Hsiang-Cheh II-73, III-356
 Huang, Hui-Yu I-327
 Huang, Kao-Lung II-672
 Huang, Po-Cheng I-33
 Huang, Qingming I-280, II-513, II-858, II-985 III-282, III-665, III-713
 Huang, Tiejun I-1
 Huang, Xiao-meng II-338
 Hui, Zhang II-197
 Huijsmans, D.P. II-229
 Hung, Chih-Hung III-697
 Hung, Shao-Shin III-542
 Huo, Longshe III-713
 Hwang, Doo Sun I-255
 Hwang, Eui-Seok I-25
 Hwang, Jenq-Neng II-947
 Hwang, Min-Cheol II-529
 Hwang, Sun-myoung I-578

 Ide, Ichiro I-247, I-650, II-489, II-657
 Ikeda, Hitoshi II-154
 Ikeda, Sei III-45
 Im, Chaetae I-564
 Im, Sung-Ho III-599
 Inokuchi, Seiji III-39
 Inoue, Kohei II-561
 Ip, Horace H.-S. II-924
 Ishikawa, Tomoya III-45
 Itakura, Fumitada I-505
 Itou, Katsunobu I-505
 Iwai, Daisuke III-39

 Jaimes, Alejandro II-154
 Jain, Lakhmi II-73
 Jang, Euee S. III-630
 Jang, Ikjin I-359
 Jang, SangSu II-1027

 Jeon, Byeungwoo II-967, III-224, III-386, III-739
 Jeong, Gu-Min III-681
 Jeong, Muh-Ho I-214, I-402
 Jeong, Mun-Ho II-736
 Jeong, Seungzoo III-23
 Ji, Zhen-Zhou III-200, III-639
 Jiang, Shuqiangu I-1
 Jin, Hai III-265
 Jin, Jesse S. I-298, II-281
 Jin, Sung Ho I-335
 Jittawiriyankoon, Chanintorn I-41
 Jo, Eun Hwan II-330, II-625
 Joseph, Vinod Cherian II-625
 Joshi, R.C. III-523
 Jou, I-Chang III-394
 Jun, Woochun II-180
 Jung, Do Joon II-641
 Jung, Doo-Hee III-681
 Jung, Junghoon III-232
 Jung, Keechul II-81, II-810
 Jung, Min-Suk II-356
 Jung, Yong Ju II-347
 Jung, Young-Kee I-222, II-866

 Kaji, Katsuhiko I-522
 Kamijyo, Shunsuke I-487
 Kamikura, Kazuto III-607
 Kanbara, Masayuki III-499
 Kandori, Keishi I-478
 Kang, Dong-Joong I-214, I-402
 Kang, Hyun-Soo III-615
 Kang, Jinyoung III-69, III-232
 Kang, Jun-Ho III-681
 Kang, Sangkyu III-69
 Kang, Sin Kuk II-810
 Kasao, Atsushi I-444
 Katayama, Norio I-650, II-489, II-770
 Kato, Noriji II-154
 Katto, Jiro I-147, III-306
 Kauff, Peter I-89
 Kawahara, Ryusuke III-53
 Kawakita, Masahiro III-31
 Ker, Jiang-Shiung I-33
 Ki, Hyunjong III-232
 Kim, Daehee III-135
 Kim, Do Young II-381
 Kim, Dong-kyoo I-556
 Kim, Doo-Hyun II-330, II-625
 Kim, Duck Hoon II-238

- Kim, Eun Yi II-810, II-1027
 Kim, Haksoo III-449
 Kim, Hang Joon II-641, II-810, II-1027
 Kim, Hansung III-87
 Kim, Heesun I-453, III-457
 Kim, Hong Kook II-381
 Kim, Hyoung-Gon II-665
 Kim, HyungJong I-564
 Kim, Il Koo III-257
 Kim, Jae-Bong II-688
 Kim, Jae-Hyun II-696
 Kim, Jae-Won III-689
 Kim, Jeong-Sik I-319
 Kim, Ji-Hee III-705
 Kim, Ji-Yeun I-255, II-786
 Kim, Jin Hak II-665
 Kim, Jin Young II-680
 Kim, Jongweon III-331, III-508, III-721
 Kim, JongWon I-139, II-364, II-373
 Kim, Kiyoung I-288, I-434
 Kim, Kyu-Tae III-508
 Kim, Kyungdeok II-762
 Kim, Manbae III-449, III-491
 Kim, Mi-Ae II-34
 Kim, Miyoung II-818
 Kim, Moon Hae II-330
 Kim, Moon-Hyun II-569
 Kim, Sang-Kyun I-255, II-786
 Kim, Sangjin III-69
 Kim, Sangkyun I-540
 Kim, Sangwook I-461, II-762
 Kim, Sehwan I-434
 Kim, Seung Man I-426
 Kim, Seung-Hwan III-159
 Kim, Seungjun III-482
 Kim, Su-Yong II-720
 Kim, Sung-Ho I-165, II-356
 Kim, Sung-Yeol I-410, III-622
 Kim, TaeYong II-146
 Kim, Tai-hoon I-548
 Kim, Whoi-Yul II-58, III-348, III-508, III-603
 Kim, Yong II-680
 Kim, Yong-Guk I-319, II-569
 Kim, Yongju I-610
 Kim, Yoon II-423
 Kimata, Hideaki III-607
 Kimber, Don I-73
 Kinoshita, Yukihiro I-626
 Kitahara, Itaru III-87
 Kitahara, Masaki III-607
 Kitasaka, Takayuki I-514
 Kiyasu, Senya II-162
 Ko, ByoungChul II-138
 Ko, Sung-Jea II-529, III-689
 Ko, You-Chang I-25
 Kobayashi, Kiichi III-31
 Kogure, Kiyoshi III-87
 Koike, Atsushi II-593
 Koike, Hideki I-97
 Komatsu, Takashi II-841
 Kong, Xiang-Wei III-426
 Kotera, Hiroaki I-190, II-794
 Koyama, Tomofumi III-15
 Kudo, Hiroaki I-532
 Kumano, Masahito III-466
 Kunichika, Yohei III-306
 Kuo, Ting-Chia III-542
 Kuo, Yau-Hwang I-33
 Kuzuoka, Hideaki I-81
 Kwak, SooYeong II-138
 Kwon, Goo-Rak III-689
 Kwon, Kee-Koo III-599
 Kwon, Young-Woo II-364

 Lai, Chun-Ming II-246
 Lai, Feipei I-114
 Lai, L. Y. III-515
 Lai, P.S. III-515
 Lai, Wei II-431
 Lam, Hong-Kwai III-314
 Lameyre, Bruno III-61
 Lan, Dong-Jun II-306
 Lay, Jose A. II-898
 Lee, Bo-Ran I-9
 Lee, ByungHo I-122
 Lee, Chen-Yi II-521
 Lee, Do Hyung II-96
 Lee, Dong Hoon I-564
 Lee, Dongwook I-139
 Lee, Doo-Soo II-569
 Lee, Eung Don II-381
 Lee, Eun-ser I-578
 Lee, GangShin I-564
 Lee, Gunhee I-556
 Lee, Hong Joo I-540
 Lee, Hyong-Woo I-25
 Lee, Hyung-Woo III-402
 Lee, Hyunju I-461
 Lee, Ivan I-131

- Lee, J.B. I-594
 Lee, Jae Gil III-290
 Lee, Jae-Yong III-689
 Lee, June-Sok III-689
 Lee, Jung-Soo II-58, III-348, III-508
 Lee, Kang-Won II-322
 Lee, Keun-Young II-967
 Lee, Kwan H. I-426
 Lee, Kyunghee II-625
 Lee, Kyung-Hoon II-529
 Lee, KyungOh I-594
 Lee, Kyu-Won I-222, II-866
 Lee, Liang-Teh II-455, III-697
 Lee, Meng-Huang I-586
 Lee, Mi Suk II-381
 Lee, Myung-Eun III-378
 Lee, Sangeun II-746
 Lee, Sangha I-556
 Lee, Sang-Kwang III-340
 Lee, Sang Uk II-238
 Lee, Seongil II-113
 Lee, Seongsoo II-537
 Lee, Seongwon III-69, III-232
 Lee, Seung-young I-548
 Lee, Seungwon I-610
 Lee, Si-Woong III-615
 Lee, Sookyong II-762
 Lee, Suh-Yin II-221, III-394
 Lee, Sun Young III-630
 Lee, Sung Yong III-290
 Lee, TaeJin I-564
 Lee, Won-Hyung II-34, II-42
 Lee, Yang-Bok II-569
 Leem, Choon Seong I-540
 Lei, Su Ian Eugene III-474
 Leo, Marco II-1019
 Leung, Yiu-Wing II-397
 Li, Chengqing III-418
 Li, Chih-Hung III-215
 Li, Hua II-553, II-617
 Li, Hui III-241
 Li, JianMin II-289
 Li, Kuan-Ching II-455, III-697
 Li, Mingjing I-198, I-344, I-352, II-213, II-553, III-111
 Li, Mingmei II-728
 Li, Peihua II-577
 Li, Shujun III-418
 Li, Weifeng I-505
 Li, Xing II-907
 Li, Yuemin II-585
 Li, Zhenyan III-575
 Li, Zhiwei I-344, II-213
 Lian, Chung-Jr II-130
 Lian, Shiguo II-65
 Liang, Yu-Ming I-206
 Liao, Hong-Yuan Mark I-206
 Liao, Xiaofei III-265
 Lie, Wen-Nung II-246
 Lim, Dong-Keun I-230
 Lim, Dong-Sun III-599
 Lim, Hyun III-378
 Lin, Chuang II-338
 Lin, Jeng-Wei I-114
 Lin, Min-Hui III-731
 Lin, Tsung-Nan II-849
 Lin, Xiaokang II-728
 Liu, Chaoqiang III-241
 Liu, Damon Shing-Min III-542
 Liu, Huayong II-197
 Liu, Kang-Yuan III-697
 Liu, Kuo-Cheng II-833
 Liu, Lin II-298
 Liu, Qiong I-73
 Liu, ShaoHui III-433, III-441, III-771
 Liu, Song I-65
 Liu, Tie-Yan III-95
 Liu, Wanquan I-385, III-79
 Liu, Yajie I-602
 Liu, Yan III-441, III-771
 Liu, Yang I-304
 Liu, Ying II-931
 Liu, Zhengkai I-344, I-352
 Liu, Zhi II-824
 Lu, Guojun II-931
 Lu, Hanqing II-704
 Lu, Hong III-143, III-575
 Lu, Lie II-890
 Lu, Xiqun I-17
 Lu, Yan II-513, III-192
 Luff, Paul I-81
 Ma, Jin-Suk III-599
 Ma, Siwei III-192
 Ma, Wei-Ying I-352, II-88, II-306, II-704, II-907, II-931, II-993, III-95
 Ma, Yu-Fei II-306
 Maddage, Namunu Chinthaka II-874
 Makita, Koji III-499
 Mao, Guojun II-172

- Mao, Ziqing II-975
 Massa, Andrea III-550
 Matsumoto, Tetsuya I-532
 Matsuo, Kenji II-593
 Mekada, Yoshito I-247
 Mi, Congjie III-143
 Mirenkov, Nikolay III-533
 Miura, Koichi II-657
 Miyahara, Sueharu II-162
 Miyajima, Chiyomi I-505
 Miyata, Kazunori I-444
 Miyazaki, Jun II-154
 Mo, Hiroshi I-650, II-489, II-770
 Modegi, Toshio III-591
 Mori, Hiroki II-736
 Mori, Kensaku I-514
 Morisawa, Keisuke II-121
 Mun, Yong-Su III-681
 Muneesawang, Paisarn I-642
 Murase, Hiroshi I-247
 Murshed, Manzur III-184
 Muto, Kenji II-405
- Nagao, Katashi I-522
 Nagashima, Shin'ichiro I-97
 Nagata, Noriko III-39
 Nakahira, Koji I-174
 Nakajima, Masayuki III-1, III-15, III-31
 Nakajima, Yasuyuki I-271, II-1
 Nakamura, Yuichi II-104
 Nakamura, Yutaka III-45
 Nakane, Kazuhiko II-27
 Nakanishi, Yasuto I-97
 Nam, Ju-Hun II-529
 Nam, Yang-Hee I-9
 Ngan, King Ngai III-167
 Nguyen, Viet Anh III-175
 Nguyen, Viet Thang III-364
 Nishibori, Kento I-532
 Nishikawa, Tatsunori I-190
 Nishino, Takanori I-505
 Nitta, Naoko I-626, II-121, II-272
- Ogawa, Masaharu II-27
 Oh, Crystal S. I-9
 Oh, Soo-Cheol I-618
 Ohnishi, Noboru I-532, II-188
 Ohsugi, Masamichi II-736
 Ohta, Yuichi II-104
 Oktavia, Vivi II-42
- Okubo, Sakae III-306
 Okumoto, Kazutaka III-39
 Otsuka, Isao II-27
- Paik, Joonki III-69, III-232
 Pan, Jeng-Shyang II-73
 Pang, Yanwei I-352
 Park, Chanmo II-373
 Park, Du-Sik II-786
 Park, Hung Kook II-601
 Park, HyeSun II-1027
 Park, Hyunwoo II-746
 Park, In Kyu I-394, II-238
 Park, Jong-Seung II-146
 Park, Jongwoon I-556
 Park, Joo-Young I-319
 Park, Kang Ryoung II-601
 Park, Sang-Hyun II-423
 Park, Sanghoon III-491
 Park, Se Hyun II-641
 Park, Seongho I-610
 Park, Soon-Young III-378
 Park, Wonbae I-359
 Park, Yongsu I-572
 Patra, Jagdish Chandra III-364
 Paul, Manoranjan III-184
 Pei, Soo-Chang I-239
 Peng, Li-Zhong II-617
 Peng, Ningsong II-824
 Peng, Wen-Hsiao II-521
 Prahmkaw, Surasee I-41
 Prameswaran, Nandan II-314
 Pu, Ruo-Nan III-143
- Qin, Lei III-127
 Qing, Laiyun II-585
 Qiu, Bin II-389
- Radhakrishnan, Regunathan II-27
 Rajan, Deepu I-65, I-182, I-470, I-634, II-993
 Ranganath, Surendra II-479
 Rao, G.S.V. Radha Krishna II-633
 Rauhala, Jukka III-298
 Rhee, Dae-Woong II-601
 Riedel, Daniel E. I-385
 Rim, Kee-Wook I-594
 Riseman, Edward II-778
 Ro, Yong Man I-335, II-347
 Romanos, Piperakis III-1

- Ryu, Jaeho III-23
 Ryu, Jeha III-482
- Saito, Suguru III-1, III-15
 Saito, Takahiro II-841
 Sakai, Shuichi II-657
 Sakauchi, Masao I-487, I-650, II-205
 Samet, Hanan II-463
 Sasmita, Lukman III-79
 Sato, Koki II-405
 Sato, Makoto III-23, III-31
 Sato, Tomokazu III-45
 Sato, Yoichi I-97
 Satoh, Shin'ichi I-650, II-205, II-489, II-657, II-770
 Schreer, Oliver I-89
 Schultz, Howard I-222, II-778
 Sebe, N. II-229
 Sénac, Christine II-882
 Seo, Dae-Wha II-712
 Seo, Yang-Seock I-255, II-786
 Shan, Shiguang II-1011
 Shao, Ling II-975
 Shenoy, Arun II-874
 Shi, Yuanchun II-754
 Shih, Arthur Chun-Chieh I-206
 Shih, Tse-Tsung II-957
 Shim, Hiuk Jae III-386, III-739
 Shimada, Kazuo II-405
 Shimizu, Satoshi III-53
 Shimojo, Shinji I-478
 Shin, Jeongho III-69, III-232
 Shin, Myong-chul I-548
 Shirai, Akihiko III-31
 Shirai, Y. I-369
 Siang, Ang Chee II-633
 Sivic, Josef II-471
 Sohn, Hyunsik III-449
 Sohn, Kwanghoon III-87
 Sohn, Sung-Hoon II-415
 Song, Bi I-344
 Song, Byung Cheol III-372
 Song, Hyunjoo II-601
 Spagnolo, Paolo II-1019
 Su, Zhou I-147
 Suenaga, Yasuhito I-514
 Sugano, Masaru I-271, II-1
 Suh, Young-Ho III-340
 Sun, Haiping II-281
 Sun, Huifang II-939
- Sun, Jinsheng II-65
 Sunahara, Hideki III-45
 Sung, Mee Young II-96, II-649
- Tai, Gwo-Chin II-50
 Takahama, Tomohisa II-794
 Takahashi, Hiroki III-1
 Takahashi, Yoshimasa II-272
 Takano, Kunihiko II-405
 Takeda, Kazuya I-505
 Takeuchi, Yoshinori I-532
 Tan, Yap-Peng III-175, III-575
 Tan, Zhang-xi II-338
 Tanaka, Hidehiko II-657
 Tanchaoren, Datchakorn II-915
 Tang, Chih-Wei III-763
 Tang, Qing II-281
 Tangerang, Ralf I-89
 Taniguchi, Rin-ichiro I-418
 Tanimoto, Masayuki I-497, III-249
 Tapaswi, Shashikala III-523
 Tehrani, Mehrdad Panahpour III-249
 Thang, Truong Cong II-347
 Tian, Q. II-229
 Tian, Qi II-19, II-264, II-314, II-479, III-566
 Tjahyadi, Ronny I-385
 Tokunaga, Takenobu III-15
 Tomobe, Hironori I-522
 Tong, Hanghang I-198, II-213, III-111
 Tong, Xiaolin III-282
 Toro, Felipe III-39
 Tsai, Piyu III-731
 Tseng, Chia-Ying III-697
 Tseng, T.C. III-515
 Tsubuku, Yosuke II-104
 Tsukada, Kiyoshi III-466
 Tung, Yi-Shin II-957
 Tyan, Hsiao-Rong I-206
- Ueda, Megumu I-418
 Ueno, H. I-369
 Ueno, Satoshi I-174
 Urahama, Kiichi II-561
- Vazhenin, Alexander III-533
 Vazhenin, Dmitry III-533
 Venkatesh, Svetha III-79
 Vetro, Anthony II-939
 Vuorimaa, Petri III-298

- Wakabayashi, Ryoji II-405
 Wan, Kongwah II-19, II-264
 Wang, Changhu II-553
 Wang, Chia-Wei III-558
 Wang, Chong II-88
 Wang, Chung-Neng III-215
 Wang, Donghui III-103
 Wang, Feng-Hsing II-73
 Wang, Haijing II-577
 Wang, Jenny R. II-314
 Wang, Jianyu II-1011
 Wang, Jinjun II-19
 Wang, Ko-Tzu II-455
 Wang, Lijun III-433
 Wang, Peng I-49
 Wang, Pi-Chung III-647
 Wang, Qinhui II-154
 Wang, Ren-Hua II-431
 Wang, Surong I-182
 Wang, Weiqiang I-263, II-609, III-127
 Wang, Wenyuan II-88
 Wang, Xin-Jing II-907
 Wang, Yanfei I-263
 Wang, Yaowei I-263
 Wang, Ye II-874
 Wang, Zhiquan II-65
 Wang, Zi-Ren III-426
 Won, Chee Sun III-583
 Won, Il-Seok II-601
 Won, Jeong-Jae I-25
 Wong, Chi-Wah III-314
 Wong, Hau San II-924
 Wong, Raymond Chi-Wing III-314
 Woo, Dong-Min I-222, II-778, II-866
 Woo, Woontack I-288, I-434
 Wu, Ja-Ling III-558
 Wu, Jiangqin I-17
 Wu, Jianhua III-673
 Wu, Jun II-289

 Xia, Tao III-241
 Xiao, Rong II-617
 Xiao, You-Neng III-143
 Xie, Dong II-975
 Xie, Jianguo III-713
 Xie, Xing II-88, II-704, II-993
 Xin, Jun II-939
 Xiong, Xuhui III-323
 Xu, Changsheng II-19, II-264, II-314, II-479, II-874, III-566

 Xu, Guangyou II-754
 Xu, Min III-566
 Xue, Xiang-Yang III-143

 Yamada, Shigeki II-728
 Yamagishi, Fuminori I-650, II-205
 Yamamoto, Daisuke I-522
 Yamashita, Jun I-81
 Yamazaki, Keiichi I-81
 Yamazawa, Kazumasa III-45
 Yan, Li II-389
 Yan, Rong II-975
 Yan, Shuicheng II-553
 Yanadume, Shinsuke I-247
 Yanagihara, Hiromasa I-271, II-1
 Yang, Chao-Tung II-455, III-697
 Yang, Chia-Lin II-957
 Yang, Jian II-890
 Yang, Jie II-824
 Yang, Shi-Qiang I-49
 Yang, Wenxian III-167
 Yao, HongXun I-304, III-433, III-441, III-771
 Yao, Min II-298
 Yashima, Yoshiyuki III-607
 Yasuda, Yasuhiko I-147
 Ye, Qixiang II-858
 Ye, Xiuzi II-298
 Yerraballi, Ramesh I-122
 Yeung, Pui Fong II-924
 Yi, Haoran I-65, I-634
 Yin, Baocai II-585
 Yin, Hao II-338
 Yip, Ben I-298
 Yokoya, Naokazu III-45, III-499
 Yoneyama, Akio II-1
 Yoo, Kil-Sang II-34
 Yoo, Kook-Yeol III-615
 Yoo, Myung-Sik III-705
 Yoon, HyoSun II-818
 Yoon, Seung-Uk I-410, III-622
 You, Shingchern D. II-505, III-151
 You, Xin-Gang III-426
 Yu, Cheng-Hsun II-505
 Yu, Chuan II-770
 Yu, Shengsheng II-439, III-323
 Yu, Xinguo II-314
 Yuan, Junsong II-479
 Yun, Il Dong II-238
 Yun, Sung-Hyun III-402

- Yun, Taesoo II-746
- Zeng, Dong II-439
- Zeng, Wei I-280, I-304, III-127
- Zhang, Baochang II-802
- Zhang, Bo II-289
- Zhang, Changshui I-198, II-213, III-111
- Zhang, Chao II-728
- Zhang, Dan III-418
- Zhang, Degan II-754
- Zhang, Dengfeng II-65
- Zhang, Dengsheng II-931
- Zhang, Fan III-755
- Zhang, Heying I-602
- Zhang, Hongbin III-755
- Zhang, Hong-Jiang I-198, I-344, II-213, II-289, II-306, II-431, II-553, II-617, II-890, II-1001, III-111
- Zhang, Hongming I-377
- Zhang, Hui I-394
- Zhang, Lei I-352
- Zhang, Liangguo II-1035
- Zhang, Ming-Ji II-609
- Zhang, Peng II-985
- Zhang, Sanyuan II-298
- Zhang, T. I-369
- Zhang, Wenli I-487
- Zhang, Xiafen I-17
- Zhang, Yuan II-513
- Zhang, Zhaoyang I-312
- Zhang, Zhengyou I-105
- Zhao, Debin I-377, II-1011, III-192
- Zhao, Feng III-119
- Zhao, Frank I-73
- Zheng, Qing-Fang II-609
- Zhong, Yuzhuo II-947
- Zhou, Jingli II-439, III-323
- Zhuang, Yueting I-17
- Zisserman, Andrew II-471