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

LNCS 3333

Advances in Multimedia Information Processing – PCM 2004

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

3 Part III

Springer

Lecture Notes in Computer Science 3333

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 III

Springer

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 3-540-23985-5 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 11363316 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** *NII/The Univ. of Tokyo*
Program Co-chairs	**Kiyoharu Aizawa** *The Univ. of Tokyo* **Yuichi Nakamura** *Kyoto Univ.* **Shin'ichi Satoh** *NII*
Poster/Demo Co-chairs	**Yoshinari Kameda** *Univ. of Tsukuba* **Takayuki Hamamoto** *Tokyo Univ. of Science*
Financial Co-chairs	**Nobuji Tetsutani** *Tokyo Denki Univ.* **Hirohisa Jozawa** *NTT Resonant*
Publicity Co-chairs	**Noboru Babaguchi** *Osaka Univ.* **Yoshiaki Shishikui** *NHK*
Publication Co-chairs	**Takeshi Naemura** *The Univ. of Tokyo* **Ichiro Ide** *Nagoya Univ.*
Registration Chair	**Ryoichi Kawada** *KDDI*
Web Chair	**Kazuya Kodama** *NII*
USA Liaison	**Tsuhan Chen** *CMU*
Korea Liaison	**Yo-Sung Ho** *K-JIST*
Advisory Committee	**Sun-Yuan Kung** *Princeton Univ.* **Hong-Jiang Zhang** *Microsoft Research Asia* **Masayuki Tanimoto** *Nagoya Univ.*

Mark Liao
Academia Sinica
Hiroshi 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.
Hiroshi 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 III

Human-Scale Virtual Reality and Interaction

Surveillance and Tracking

Image Analysis (III)

Compression (II)

Streaming (II)

Watermarking (II)

Content Production (II)

Applications (II)

Multimedia Analysis

Compression (III)

Watermarking (III)

Table of Contents, Part I

Volume I

Art

Network (I)

Sports (I)

Immersive Conferencing: Novel Interfaces and Paradigms for Remote Collaboration

Network (II)

Image Retrieval

Image Analysis (I)

Face, Gesture, and Behavior (I)

Virtual Reality and Computer Graphics

Content Production (I)

Intelligent Media Integration for Social Information Infrastructure

Approaches or Methods of Security Engineering

Multimedia Servers

Video Retrieval

Table of Contents, Part II

Volume II

Application of Video Browsing to Consumer Video Browsing Products

Watermarking (I)

User Interface (I)

Content-Based Image Retrieval

Sports (II)

Network (III)

Streaming (I)

Visual Content Mining in Multimedia Documents

Compression (I)

Face, Gesture, and Behavior (II)

Applications (I)

User Interface (II)

Image Analysis (II)

Audio Analysis

Moving from Content to Concept-Based Image/Video Retrieval

H.264

Face, Gesture, and Behavior (III)

WARAJI: Foot-Driven Navigation Interfaces for Virtual Reality Applications

Salvador Barrera, Piperakis Romanos, Suguru Saito,
Hiroki Takahashi, and Masayuki Nakajima

Nakajima Laboratory, Tokyo Institute of Technology,
W8-64/70 2-12-1 Ookayama Meguro-ku 152-8552 Japan
{salvador,romax,suguru,rocky,nakajima}@img.cs.titech.ac.jp

Abstract. Presently Technological limitations on current interfaces have made researchers to develop new devices to interact with objects in the virtual environment. The goal of this project is to develop and build a hands-free navigation system to be integrated into virtual environments. One of the most important fields in virtual realty (VR) research, is the development of systems that allow the user to interface with the virtual environment. The most intuitive method for moving through a virtual landscape is by walking. Systems ranging from different platforms have already been implemented to produce virtual walking; however, these systems have been designed primarily for use with head mounted display systems. We believe that hands-free navigation, unlike the majority of navigation techniques based on hand motions, has the greatest potential for maximizing the interactivity of virtual environments, due to more direct motion of the feet.

1 Overview

Electronic sensors have been incorporated into footwear for several different applications over the last years. Employing force-sensing resistor arrays or capacitive sensing, insoles with very dense pressure sampling have been developed for research. As sensors and associated processing systems decrease in cost and bulk, they also begin to adorn athletic footwear. Examples are pressure-sensing insole for golfers to improve their balance during a swing. Although most interfaces for virtual reality applications concentrate on the hands, fingers, and head, some have been extended to the feet. as an example, the "Fantastic Phantom Slipper", where a pair of infrared-emitting shoes are tracked over a limited area and haptic feedback applied by driving vibrators in the sole [6]. Focusing on a different approach for virtual reality input devices, "WARAJI I" was made, WARAJI(Walking, Running and Jumping Interface), had pressure sensors on each foot sole as shown in Figure 1.

Users feet can be used for input basic operations giving freedom to other parts of the body. This version was rather primitive and was based on detecting the pressure orientation of the sole. "WARAJI I" was implemented to cope with the flaws in the design. More specifically the first prototype suffered from

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 1–7, 2004.

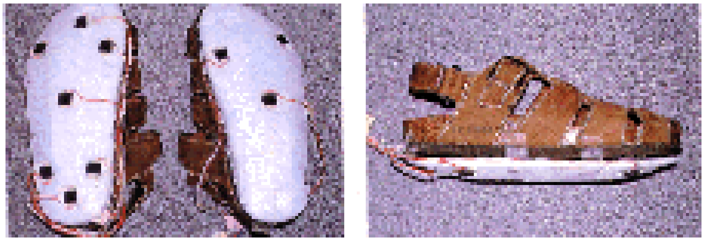


Fig. 1. “WARAJI” First Input Device based on pressure sensors

lack of sensitivity in backward weight shifts. After many experimental results we designed a new version of the foot input device known as “WARAJI II” as shown in Figure 2. This version of foot input was based on detecting the orientation of the sole, converting it to a directional signal [1]. This implementation had rotary encoder sensors on a sole.

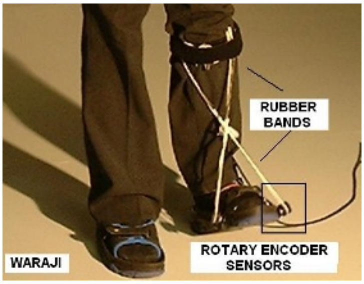


Fig. 2. “WARAJI II” Foot input device based on rotary encoder sensors

“WARAJI II” was based on rotary encoder sensors. The sensors are mounted on the sides of the sandal and are attached to a Velcro strap located around the knee via rubber bands as can be seen in Figure 2. This simple device detect ankle movements. These sensors rotate according to the amount and direction of the movement of the foot. The sensor then collects analog information through the user’s movements from the legs. The emphasis is in the processing and collection of motion and position data that the sensors accumulate. Rotary encoders serve as measuring sensors for rotary motion, and for linear motion when used in conjunction with mechanical measuring standards. These sensors are made on the basis of magneto-induction transducers and are used in the control system where determination of rotation angles, number of revolution, and speed or rotation. The sensors measure the rotary motion of the user’s feet and are part of “WARAJI II”. Therefore the sensors and the rest of the system work together [3].

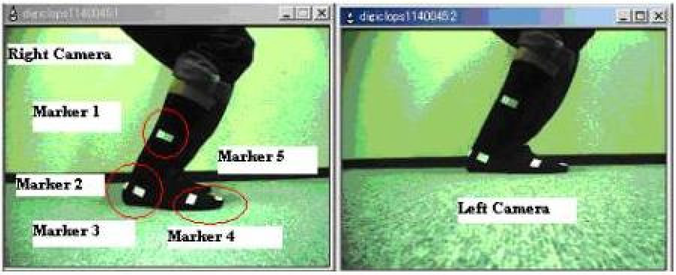


Fig. 3. Markers positions of the foot and the segmentation planar model

“WARAJI III” used stereoscopic images from the foot of the human body and were used to estimate the foot motion in three-dimensional space. The basic approach is to store a number of 2D views of the human feet in a variety of different configurations. Markers were attached to the surface of the foot based on the known location of the forefoot, variants include the case of camera viewing the same tracking the foot configuration and pose over time from video input as shown in Figure 3.

Finally, “WARAJI IV”. A simple device using acceleration sensors to detect ankle movements within the virtual environment. The acceleration sensors are attached to the foot and detect movement based on direction for three different angles as shown in Figure 4. The forces of acceleration move the piezoelectric seismic mass, thereby causing strains to it, which generates the voltage [4]. The sensors measure acceleration in three directions x, y and z. sensing the detectable ankle motions for the foot movement. “WARAJI IV” permits users freedom for changing motion directions naturally. This allows users to input two or more operations simultaneously. Examples include pointing out an object and changing position at the same time. Users can express where they want to go or what they want to do trough natural movements. This also allows the user to move, jump or walk without making any step or hand movement. Such an interface presents a series of design choices, centered on the user control and number of degrees of freedom to be presented. We have set out to make these interface as “natural” as possible. This experimentation could prove beneficial in future virtual gaming [5]. Validation of our approach is given by discussion and illustration of some results.

2 WARAJI II Versus WARAJI IV

“Waraji II” consists of two rotary encoder sensors attached to a sandal for detecting motion. The sensors are connected to a strap right below the knee with rubber bands. These sensors rotate according to the amount and the direction of the foot. The angle of the ankle is detected and translated to an electrical signal, the PC detects changes in voltage and calculates the direction data and inputs it into graphic system. Since the level of the voltage that the interface outputs is

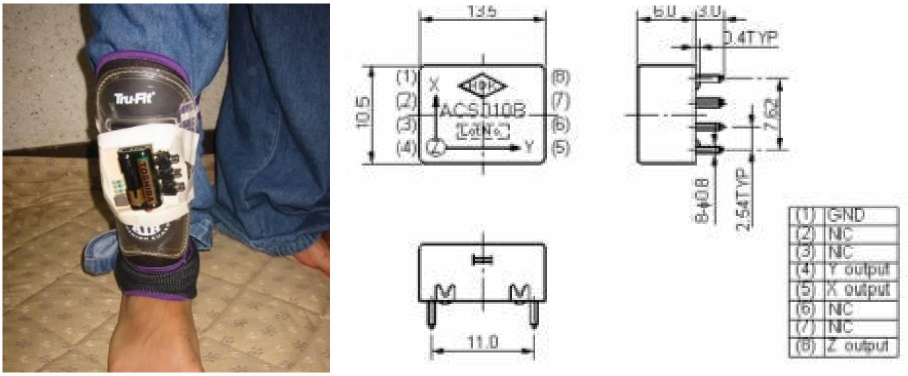


Fig. 4. "WARAJI IV" Foot motion sensing device based on acceleration sensors

in an analog form, we need a sampling process to convert it into a digital signal that our PC can manipulate. "Waraji II" is connected to an A/D conversion board which takes care of the conversion in a rate that provides the user with high play ability and unnoticeable response times. The PC detects changes in voltage, calculates and sends direction data to the graphic system according to increases or decreases in voltage sending direction data to the graphic system such as Figure 5. In this application, because we choose a game which only requires two dimensional input, we use only two sensor values. However for other cases which require three dimensional values, the algorithm could not be extended to incorporate vertical, horizontal and rotationally values. Starting from this view point was necessary to upgrade the version using acceleration sensors. These sensors permit users to sense foot motion in 3 different dimensional values.

2.1 WARAJI IV

"WARAJI IV" consists of some acceleration sensors attached to the human leg for detecting motion. The sensors are connected with some rubber bands directly below the knee [1]. The acceleration sensors sense foot motions and translate that action into movement. Specifically, the acceleration sensors measure the acceleration, direction and amount of the foot's motion as shown in Figure 4. According to the amount and the direction of the foot, the angle of the ankle is detected and translated to an electrical signal. Since the level of voltage of the interface is in an analog form, we need a sampling process to convert it into a digital signal that our PC can manipulate. "WARAJI IV" is connected to an A/D conversion board which takes care of the conversion in a rate that provides the user with high play ability and unnoticeable response times. Afterwards the PC detects changes in voltage, calculates and sends direction data to the graphic system, according to increases or decreases in voltage.

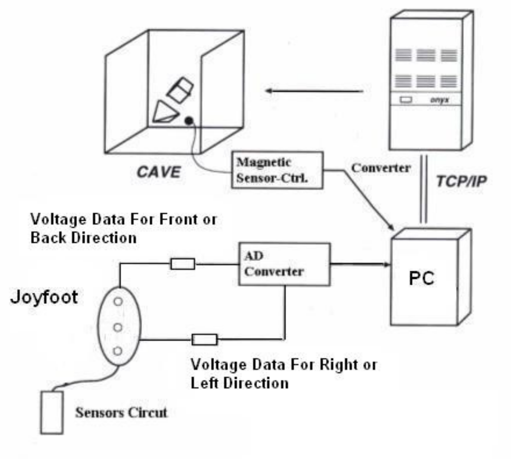

CAVE
onyx
Magnetic Sensor-Ctrl.
Converter
TCP/IP
Voltage Data For Front or Back Direction
AD Converter
PC
Joyfoot
Voltage Data For Right or Left Direction
Sensors Circut

Fig. 5. Overview of the System Architecture

3 Hardware System Architecture

Since various scanning modes are being investigated we use a stereo viewing system for displaying., which results in a number of systems incompatible to one another. We address the problem of the interconnection of such a device through standard conversions by a signal processing approach,we used stereo viewing system for displaying. Namely the model of a universal standard converter, which is based on a layered functional architecture. The concept of a virtual standard is introduced for stereoscopic signals. When this machine receives direction data, it redraws a picture according to the data. In summary, the system receives direction data from the “WARAJI” unit, reconstructs the scene picture and transmits it into the projector as shown in Figure 5.

4 Data Acquisition Technique

Before using the “WARAJI” calibration is needed, this requires five key-points: Center (C) to serve as the neutral position, Front (F), Left (L), Back (B) and Right (R). The vectors CR, CF, CL and CB divide the sensor plane to four parts, which are mapped to the four quarters of XY. Each of these vectors is moved, rotated, sheared and re-scaled to coincide with the vectors x, y, -x, -y of the target system. The transformation algorithm takes into account the possibility that the key points form a left-handed coordinate system, and manipulates the values. These points form a region of all the possible sensing values for each angle of the user’s foot. This is called The user’s sensing plane. Vectors from the Center point to the four other points are used to decompose arbitrary directions to a regular coordinate plane as shown in Figure 6.

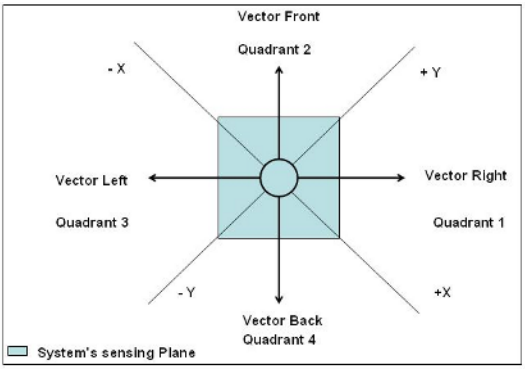


Fig. 6. System's sensing plane

5 Experimental Results and Conclusions

A number of design choices were presented and centered for the user control that can be objectively used in many Virtual Reality applications. Compared with the traditional input devices such as keyboard and mouse, "WARAJI" opens a new wearable user interface technology with motion detected from the foot. Giving freedom to other parts of the body, users can interact easily in the virtual world, making hands free from motion operation. Foot operations are very useful for moving naturally around the virtual world. As it is shown in Table 1, "WARAJI I" presented some problems such as Balance, Critical Mass and Sensitivity since the pressure sensors were concentrated in the center of WARAJI. This version could not succeed since data were received from the device when there was no movement. Focus on that view point "WARAJI II" was made demonstrating the power of body sensing by giving freedom to our hands and making games interactive to users, we use only two sensor values. However for other cases which require three dimensional values, the algorithm could not be extended to incorporate vertical, horizontal and rotationally values, same as more freedom. "WARAJI III" was developed using a stereoscopic camera. We took stereoscopic images from the foot, located the key points and used these to estimate the foot configuration and pose in three-dimensional space. Users were able to walk in the virtual environment without any electronic equipment attached but not in real time. Since part of this version work only for still images we could not succeed. The algorithm and the camera calibration need to be improved. Since new technological advantages we proceed to create "WARAJI IV" This version gives the user more freedom, as well as, measurement results with horizontal, vertical and gravity ratings of motions. The system is distributed and depends on how much the person can move his/her foot at the time they wear the device. The output is generated and changed at the very moment when the acceleration is applied. In combination with the previous versions, users can also travel in directions that were originally directly behind them when they faced the front

Table 1. WARAJI:Merit and Demerit

WARAJI	Version I	Version II	Version III	Version IV
Good for real time	O	O	X	O
No weight limitation	X	O	O	O
Freedom of movement	O	X	O	O
3 Degees of freedom	X	X	X	O

wall of Cave by first turning to the body either the right or left, also is not weight dependent. We have observed that user's need time to adjust to this distorted spatial mapping, but can at least navigate in any direction. However, we have not yet attempted to quantify the effect of this auto rotation technique on a user's sense of spatial relations.

References

[1] S. Barrera, H. Takahashi, M. Nakajima, A new Interface for the Virtual World Foot Motion Sensing Input Device, Conference Abstracts and Applications, SIGGRAPH 2002 Computer Graphics Annual Conference Series, San Antonio: 141

[2] S. Barrera, P. Romanos, H. Takahashi, S. Saito, M. Nakajima, Real Time Detection Interface For Walking on CAVE, Proceedings Computer Graphics International, CGI 2003, Proc of IEEE 2003 Tokyo, Japan, July 9–11, 2003

[3] S. Barrera, H. Takahashi, M. Nakajima, "Foot-Driven Navigation Interface For A Virtual Landscape Walking Input Device", Beyond Wand and Glove Based Interaction – VR Chicago 2004, Proceedings IEEE VR 2004, Chicago, USA, March 28, 2004

[4] S. Barrera, H. Takahashi, M. Nakajima, Hands-free navigation methods for moving through a virtual landscape, Computer Graphics International, CGI 2004, Proc of IEEE 2004 Crete, Greece, June 16–18, 2004

[5] S. Barrera, H. Takahashi, M. Nakajima, "Joyfoot‘s Cyber System: A Virtual Landscape Walking Interface Device for Virtual Reality Applications", Proceedings of the 2004 International Conference on Cyberworlds (CW2004) Proceedings IEEE, Computer Society Press. Cyber Worlds 2004, Tokyo, Japan, November 18–20, 2004

[6] M. Sato, Y. Kume and M. Kusahara, Foot Interface: Fantastic Phantom Slipper, SIGGRAPH98 – 25th International Conference on Computer Graphics and Interactive Techniques, Orlando, 1998

Time Space Interface Using DV (Digital Video) and GPS (Global Positioning System) Technology – A Study with an Art Project "Field-Work@Alsace"

Masaki Fujihata

Department of Inter Media Art, Fine Arts,
Tokyo National University of Fine Arts and Music
5000 Komonma, Toride, Ibaragi, Japan
masaki@fujihata.jp
http://www.fujihata.jp, http://www.ima.fa.geidai.ac.jp/~masaki

Abstract. This paper is summarizing a media art project from the artistic idea to technological realization, and tried to conclude the important aspects of media art from both side. Auther is an artist had been working as a pioneer in this area from computer graphics in early eighties until interactive media art. The project "field-works@Alsace" is a truly successful example in a context of "Interactive Cinema." Especially with the CAVE environment, its 3 dimensional space made possible to tactile the content of the piece very well. Originally it was not planned to use the CAVE, but the CAVE is the ultimate system for this project.

1 Preface

As a media artist, media technology is a tool to create a new medium, which enables to realize an artistic vision into real production. The following example, "Field-work@Alsace" is the actual art project which was co-produced with ZKM(Center for Art and Media Technology), Karlsruhe, Germany and myself. Within the artistic side, the aim of this project is collecting people's voice who is living near the border between Germany and France by the video interview with GPS. Technical aspect of this project is a sensor fusion of positioning and orientation (directional) data with moving images. By establishing this fusion, at the final image at the CAVE [1,2,3] screen, the video image is projected on the virtual screen at the place where the image was shot and is moving according to the movement of the camera, which can create an illusion, for example, as far as the screen moves tilt or pitch the horizontal line in the video image is not moving at the center as it is.

2 Location and Image

The fundamental idea for this series of project titled "Field-Works" is combine two different information, the position data and video images into one

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 8–14, 2004.

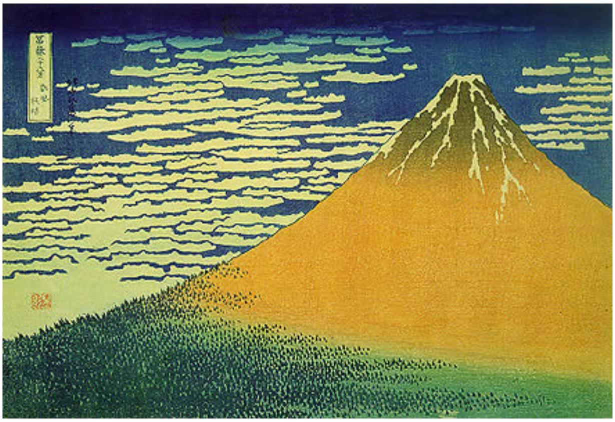


Fig. 1. "Miage no Fuji" Katsushika Hokusai, wood block printing The height of the Mt.Fuji is about 1.7 times higher than the real.

cyberspace. In 1992, I experimented with GPS (Global Positioning system) for recording the series of position data, longitude, latitude and altitude while I was climbing up the Mt. Fuji that is the highest mountain in Japan. The original idea for collecting GPS data was to deform the shape of Mt.Fuji to fit the impression of climber. "Deformation" is one of the traditional methods for paintings. In history, Mt. Fuji had been deformed many times by many famous painters and wood block printing maker in the past. (See. Hokusai's famous Mt. Fuji picture (Fig. 1)) After the real experience of climbing the Mt. Fuji, the resulting image (Fig. 2) is made and is showing an exploded shape of Mt.Fuji which was deformed by my slowness for climbing up that is caused by tiredness of my foot. This exploded shape was made by scaling the section of each altitude according to my speed of that altitude; speed was calculated from GPS data.

At the same time, I also recorded whole sequences with video-camcorder, Sony Hi-8 at that time. For the show in 1994, the special archiving software was programmed for locating video images to the GPS data. (Programmed by Nobuya Suzuki with the workstation.) It is a good example that is showing a new way of handling video sequences with its location. Location can be used as a tag for searching a video sequence. Each yellow tag on the 3D GPS line that can be clicked by mouse is designed for activating the movie on the desktop window. (See Desktop image from Irix workstation (Fig. 3))

After the year 1994 I had been stopped to continue this project, from the year 2000 I started again the similar project titled "Field-Works". The year 2000 is a mark able year in the field of GPS, because president Clinton stopped the scramble from consumer use signal of GPS from 2nd of May, which gives us completely same accuracy with military use the accuracy is plus minus 5m.

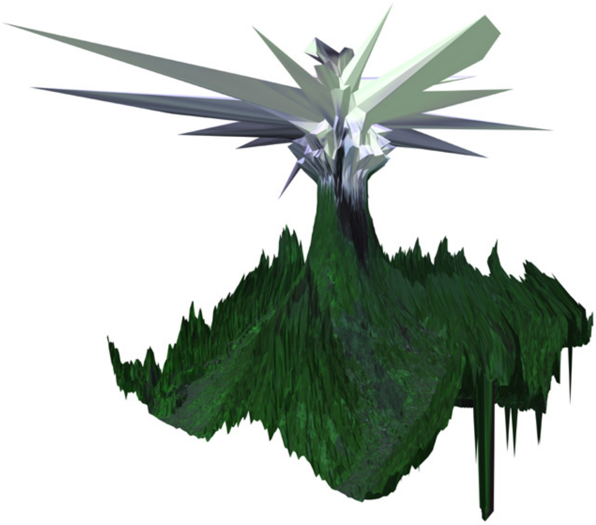


Fig. 2. “Impressing Velocity” 1992-94 Masaki Fujihata Resulting; image deformed shape was made with the climbing up data.

3 Starting Field-Works Project [4]

After two years experimental projects realization from the year 2000, I started new project that targeted the border between Germany and France, for the coming exhibition “Future Cinema” which would be organized by ZKM in 2002 [5]. The border exists an abstract conception, but once it is activated it changes the status and is possible to kill people. However the border line is not visible in real location even visible on the map, the idea for the project was coming to my mind the border can be visualize by tracing with my foot and GPS. Of course these GPS lines can contain video sequences of the interviews with local living people near the border.

I targeted the area called Alsace. Strasburg is the main capital of this region and now is famous for the center for the European Union. Alsace is not France and not Germany; even the political situation of this area had been changed to French or to German several times. Still they are independent from others and still some of them can speak their own language “Alsacisch.” In this area, we can here German speaking, French speaking, and Alsacisch speaking which can make a border between different languages, but in real most of the living people can speak all of them. It’s a multi-lingual situation and for the people who can speak three languages the border is meaningless, it is internationally borderless. The real border is more complex.

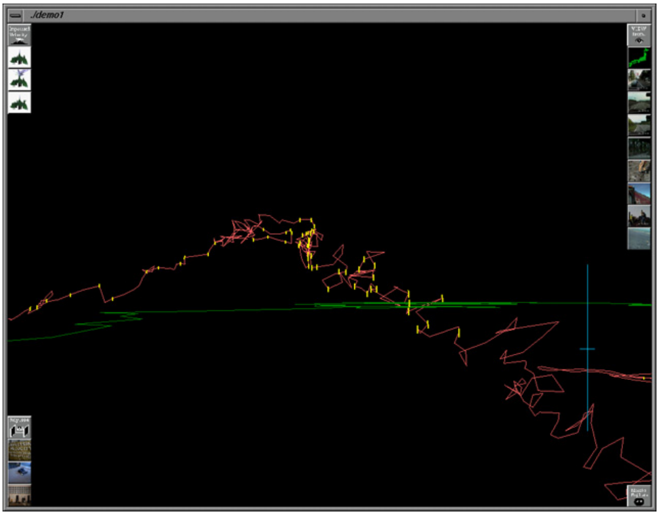


Fig. 3. "Impressing Velocity" 1992-94 Masaki Fujihata Computer display capture image; jamming of lines are the GPS data and yellow tiny vertical lines are the hot spots for the movie.

One clear border I found while in the process of the project is IT border. Each mobile phone company should account the transactions when the route was changed at the border that reflects the charge. Once you cross the border from Germany to France, at the each time the mobile is annoying with sound, which tells you that the welcome message was received and new roaming service started even when we are speaking same language before and after crossing the border.

4 Position and Orientation

Real space where we are living in is three dimension and we are passing through the time. Time cannot be experienced from backwards. Photography made possible to cut off two-dimensional image from three-dimensional space and Cinematography made possible to record series of two-dimensional images from three-dimensional space and time. It was believed that the resulting image cannot contain of its location and is the main characteristics of the image generation. Attracting point of photography is photo image may invite us to start to imagine the place where the image was fixed. Our imagination can bring us to the place where we do not know, we also do not care even the place was exist when the photo was recorded. On the other hand, when the photography was used for documenting the fact, a photo need to be attached with a written text which documenting the happening, location, and relation. The position and orientation data can make photography more valuable media.

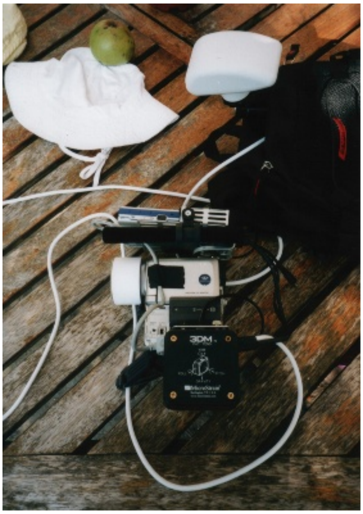


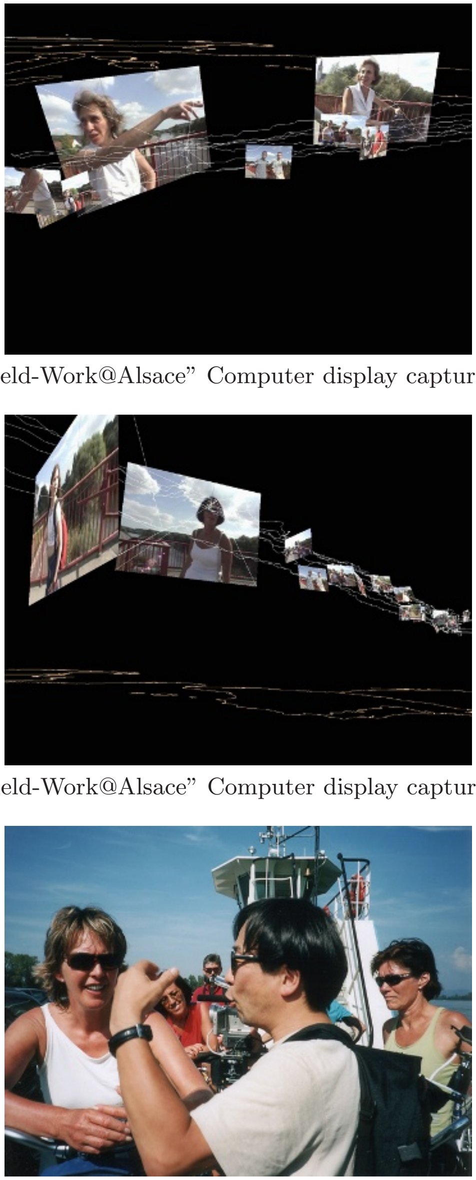
Fig. 4. Equipments; DV camera, PDA, GPS, and 3D strain

In the case of “Field-works” project, the location is captured by using GPS even the video camera was not running, for example while in transporting from main camp to the real location, the position data have been captured. Technically speaking, it is quite important to record position data without video sequences, because these lines can be used to figure out the shape of that region near the video image was shot. In this case, GPS data recording is dominant rather than video images.

Another important data for the final production is the orientation data that is showing the directional data on which direction the camera was targeting. It is used for commanding the direction of the screen in the final cyber space. “3DM” is the name of the sensor made by 3D strain company. (See Fig. 4) The movement of the screen is showing the movement of the camera which indicating the intension of the cameraman what he was willing to shoot. At the show in CAVE the user can hear the interview itself and see the faces and some more sceneries and also user can read the movement of the author as a cameraman.

5 Reality with the Screens in the Screen

Cinema was invented in early 20 centuries. Shooting a moving image and seeing a moving image is quite new experience for the human’s history. Shoot by camera and then projected on the screen is the typical scheme for the cinematic system. Within this system the camera can move but the screen is not moving. Even for the television system, viewer had been sitting in front of the television set. For the coming situation of media so called multi-media, new-media, or networked media, this common sense will be alternated by the new medium which is not



(a): “Field-Work@Alsace” Computer display capture image

(b): “Field-Work@Alsace” Computer display capture image

(c): Snap shot on the ferry boat on the Rhein river

Fig. 5. (a) and (b) are taken from computer screen. Thin white lines are made from GPS data which shows the movement of the cameraman, and the rectangle frame is texture mapped with video streaming. While the video is running on it, the rectangle frame is also move according to the orientation data. For example when the camera move to right, then the rectangle is also move to right. Each video was edited in post process, each video’s duration is about 30 sec. to 90 sec. and the maximum three videos are running simultaneously.

yet invented and will make a new common sense that the screen will fade away or screen can contain screens or screen can go anywhere. The CAVE, originally proposed EVL at Illinois's university is a good example for this topic where people can forget about the real vinyl screen, and then the viewer can focus the virtual object in cyber space.

The interaction with images and real 3D space movement the new type of reality might come up to the user. It is not a kind of virtual reality also not a kind of mixed reality, but the user's imagination can be convinced to construct a spatial reality by combining 3 dimensional movements in cyber space and with many numbers of the fragmented video images which had shot in real space. The importance is how to kick and start user's imagination for creating and connecting several different realities into one. The interface, in this case whole design of the medium, should be designed very carefully and centered the user's behavior. Designer should learn the user's process of learning the language of the interface.

In my case, as an artist, who designed whole process of this project and realized it by himself until exhibit it. When at the starting point, no one could imagine how the result is becoming; only the artist could recognize and was possible to start the production. Each tiny movement of the screen in cyber space was calculated when it was recorded on site according to the consciousness of the artist for the final production. I need a new design of the medium and then I constructed and used it. Technically speaking in the engineering side, it is possible to create any type of new medium or combine several different mediums into one or even improve the pre-exist medium, but it is not usable for the artistic production when it could not give the artist to imagine the valuable aspect of the medium itself.

References

[1] Carolina Cruz-Neira and Daniel J.Sandin and Thomas A.DeFanti, "Surround-Screen Projection-Based Virtual Reality – The Design and Implementation of the CAVE", Proc. ACM SIGGRAPH 93, pp.135–142 (1993)
[2] Ars Electronica Center, Linz, Austria, *http://www.aec.at*
[3] National Museum of Emerging Science and Innovation. *http://www.miraikan.jst.go.jp*
[4] Field-works project page: *http://www.field-works.net*
[5] "Field-works@Alsace" at the exhibition "Future-Cinema" at ZKM Karlsruhe *http://www.zkm.de/futurecinema/fujihata_werk_e.html*

Action Generation from Natural Language

Satoshi Funatsu, Tomofumi Koyama, Suguru Saito,
Takenobu Tokunaga, and Masayuki Nakajima

Department of Computer Science
Tokyo Institute of Technology
Tokyo Meguro Oookayama 2-12-1, Japan 152-8552
{funa@img,tomoshit@img,suguru@img,take@cl,nakajima@img}.cs.titech.ac.jp

Abstract. When building a virtual agent system equipped with a natural language command interface, the interpretation of sentences into action commands that will drive the animation is crucial. In our system, these action commands are transformed a series of spatial and action constraint symbols. Using these symbols, the system can decide on the most plausible motion for the agent.

1 Introduction

In recent years, there has been a considerable interest in simulating human behavior in both real and virtual world situated scenarios [1,2,3]. If simulated agents or robots could understand and carry out instructions expressed in natural language, they could vastly improve their utility and extend their area of application. However in general, linguistic expressions have ambiguity and vagueness. It is thus often hard to resolve the ambiguity in an automatic manner. For example, in the sentence "stand in front of the table", the word "front"has vagueness. Finding an exact such location for the agent is difficult even though a human person could perform the same task instinctively. In this work, we are focusing on the problem of natural language command driven motion generation. In particular, our aim is twofold. First, to express the constraints specified explicitly by the user, with those implied by the virtual character's body and the surrounding environment, into a uniform representation; and second to develop a system that uses this representation in order to generate smooth agent animation consistent with all the constraints.

2 System Overview

Figure 1 shows a screen shot of the system. There are two agents and several objects (colored balls and tables) in a virtual world. Using speech, a user can command the agents to manipulate the objects. The current system accepts simple Japanese utterances, such as "Tsukue no mae ni ike." (Walk to the table) or "Motto"(Further). The agent's behavior and the subsequent changes in the virtual world are displayed to the user as a three-dimensional animation.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 15–22, 2004.

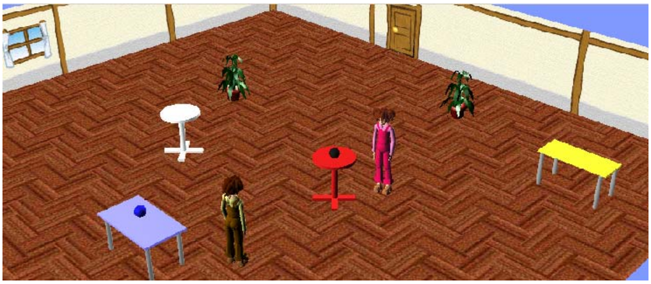


Fig. 1. Screen shot of the agent system

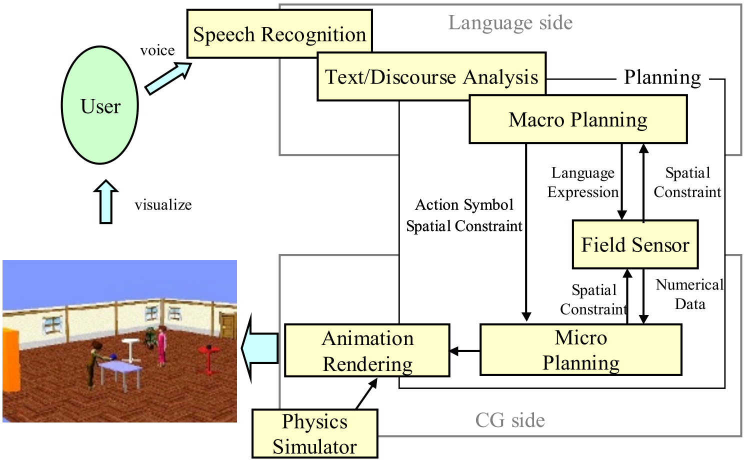


Fig. 2. System architecture

Figure 2 illustrates the architecture of the system. The speech recognition module receives the user's speech input and translates it to a sequence of words. The text/discourse analysis module analyzes the word sequence to extract a case frame,extracts the user's goal and passes it over to the planning modules which build a plan to generate the appropriate animation. In other words, the planning modules translate the user's goal into animation data. However, the properties of these two are very different and straightforward translation is rather difficult. The user's goal is represented in term of symbols, while the animation data is a sequence of numeric values. We separate planning stage into two stages – macro and micro planning to account for the differences in representation.

During the macro planning, the planner needs to know the qualitative properties of the involved objects, that depends on their size, location and so on. For example, in order to pick up a ball, the agent first needs to move to a location from which he can reach the ball. In this planning process, the distance between the ball and the agent needs to be calculated. This sort of information is represented in terms of virtual space coordinates and is handled by the micro planner.

To interface the macro and micro planning, we use the SPACE object. This object is provided by the field sensor module and is capable of representing a location in the virtual space, in a bilateral character; symbolic and numeric.

The micro planning module is responsible for the details of the agent's motion. It receives action and spatial constraints from the macro planner in the form of SPACE objects, and composes a keyframe sequence which satisfies both types of constraints by using a motion database included in the module. These spatial constraints are changed into numeric values by the field sensor module, which are in turn used by the micro planning module for the evaluation of candidate motions. The keyframe sequence is passed to the animation rendering module, and an animation that satisfies both the user's requests and the current environment configuration is produced.

3 SPACE Object

When we generate animation from instructions expressed in natural language, it is necessary to deal with spatial constraints. For example, the instruction "Tsukue no migi wo tootte tana no mae ni itte"(Go to the front of the shelf via the right of the desk.) contains two spatial constraints. First, the agent has to move to the area referred to as "tsukue no migi"(right of the desk). Next, it has to reach the destination referred to as "tana no mae"(front of the shelf). However these spatial expressions are vague and do not point to exact positions necessary for generating a specific action. To overcome this problem we introduced a data object called "SPACE", which bridges the gap between symbolic expressions of spatial relations and their plausibility in virtual space [4].

Every SPACE object has a potential function. This function quantifies the concept of a spatial constraint. Therefore SPACE object deals both with environmental and the character's body constraints within the same framework. The potential functions are designed to conform to two conditions. Differentiability throughout their domain, to be able to find the maximum using Steepest Descent and bounded between 0 and 1, to interpret the result of logical AND and NOT operations on SPACEs as a plausibility. By operating on SPACE objects, virtual agents can interpret complicated spatial expressions and evaluate constraints in complicated environment configurations.

Figures 3 and 4 show isosurfaces of the potential field of SPACE objects corresponding to ‘‘right" and ‘‘by." The potential of directional spatial nouns is defined to decrease in relation to the distance from the characteristic basic semiaxis. The potential of distance-spatial nouns is based on the distance from the reference object's convex hull.

4 Motion Generation in the Virtual World

The micro planning module generates motion sequenses by using a motion database called MotionGraph [5]. This is an extension of the keyframing method used to generate natural motions and several similar approaches based on the

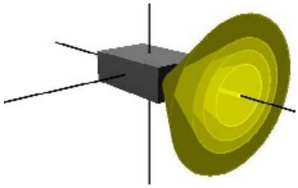


Fig. 3. Isosurface of potential field for "*m*igi(right) SPACE"

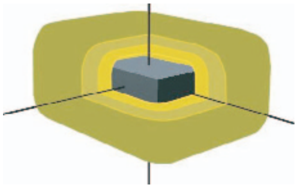


Fig. 4. Isosurface of potential field for "*s*oba(by) SPACE"

same concept can be found in the literature [6,7]. MotionGraph is a directed graph where each node represents a given posture of a character and each link, properties of the transition between consecutive keyframes such as the translation vector, global rotation, duration of the transition and a pointer to a default next node, as shown in Figure 5.

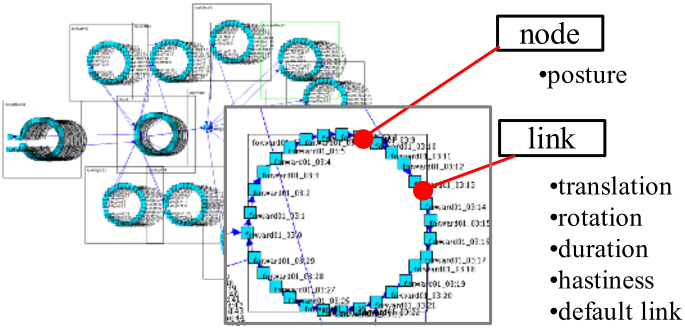


Fig. 5. MotionGraph

When a simple action order is given, the system first determines a valid goal node. Then a list of all the nodes residing on the shortest graph path from the current to the goal is used to compose the animation keyframes. Furthermore, we can use the "hastiness" property of the graph links that allows us to further control the selected path. When a complex action command is given, it is translated to a node group. In the case of locomotion, the latter is used.

The micro planning module searches the best graph path in a node group subject to optimizing the cost of the motion and spatial constraints. These constraints are derived both from the issued command and the environment's obstacle configuration. As mentioned, these are dealt with using the potential functions of SPACE objects. To find an appropriate path on a MotionGraph under constraints we need to make a temporal tree of candidate paths and to project it on the potential field composed by the functions of the involved SPACE objects. This task involves the following steps.

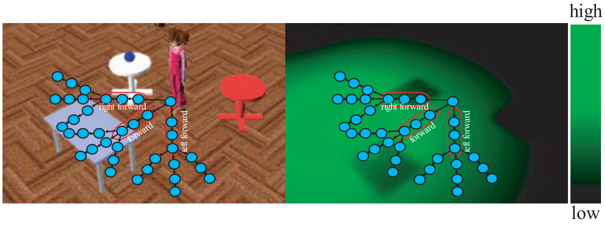


Fig. 6. Mapping of MotionGraph to potential field

1. set the current node on the graph that corresponds to the agent's current state as the root of the temporal tree.
2. follow links originating from that node in search of candidate transition nodes
3. add each of these nodes to the temporal path tree and create the appropriate path branches.
4. if the elapsed time from the root to a leaf node exceeds a certain limit, stop extending this branch.
5. otherwise, perform 2nd step for each link on the branch reccursively.
6. when every path has been extended to its limit, map the tree to virtual world coordinates as in Figure 6.

Every branch corresponds to a possible motion sequence. The most suitable one is chosen to be carried out. This process is repeated at short intervals. The choice of the most appropriate path is controlled by a function U defined as in Equation (1). The path with the highest overall evaluation of the function is deemed the most suitable one.

In Equation (2), (3), and (4), where $\boldsymbol{p}$ is the position at node i , E is the potential value at $\boldsymbol{p}$, $t(i, i+1)$ is time elasped between node i and $i+1$,r is front direction of the agent at node i, U_1 is the value of a quadrature by parts for the potential field along a locomotion path. Since we require agent moves in a natural, smooth motion, U_2 and U_3 evaluate second differential of of the agent's position p and direction r respectively.

$$U = U_1 * \alpha + U_2 * \beta + U_3 * \gamma \tag{1}$$

$$U_1 = \sum_{i=1}^{k-1} \frac{E(\boldsymbol{p}(i)) + E(\boldsymbol{p}(i+1))}{2} * t(i, i+1) \tag{2}$$

$$U_2 = -\sum_{i=1}^{k-2} \left| \frac{\boldsymbol{p}(i+1) - \boldsymbol{p}(i)}{t(i, i+1)} - \frac{\boldsymbol{p}(i+2) - \boldsymbol{p}(i+1)}{t(i+1, i+2)} \right| \tag{3}$$

$$U_3 = -\sum_{i=1}^{k-2} \left| \frac{r(i+1) - r(i)}{t(i, i+1)} - \frac{r(i+2) - r(i+1)}{t(i+1, i+2)} \right| \tag{4}$$

5 Result

Using our character agent system, we generated a variety of different user command-driven animations. Here we show locomotion example sequence. Figure 7 shows the initial state in which a character named Natchan stands still inside a room. Figure 8 shows the result when "Natchan ha aoi tukue no mae ni ike" (Natchan, go in front of the blue table.) is given as a user's input. The gradation coloring of the floor shows the value of the potential field and the line signifies the generated trajectory. Figure 9 is the result of issuing the "kuroi tama no migi ni ike" (Go to the right of the black ball.) command. Figure 10 shows the next state when "hidari no shiroi tukue wo miro" (Look at the white table of the left.) is given. Figures 11 and 12 show the result when "migi no aoi tama wo kiiroi tukue ni oke" (Pick up the blue ball of the right and put it on the yellow table.)

In the presented scenario, the commands were issued one after another.Note that the completion of some commands depends on the current the status of the agent. The macro planner resolved the goal and identified the objects involved and the micro planner generated the details of the actual animation. The results show that the generated plans are performed successfully using the macro and micro planning modules.

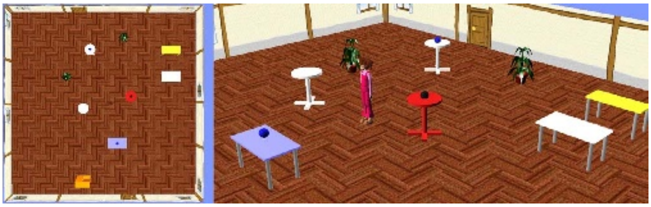


Fig. 7. Initial state

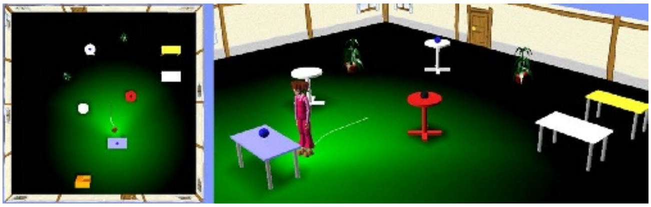


Fig. 8. "Natchan ha aoi tukue no mae ni ike" (Natchan, go in front of the blue table.)

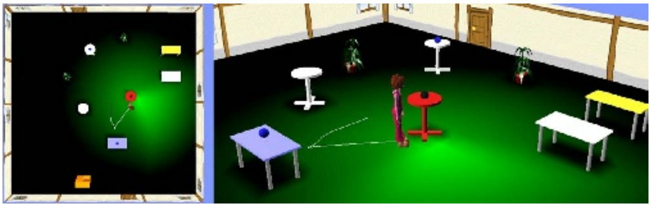


Fig. 9. "Kuroi tama no migi ni ike" (Go to the right of the black ball.)

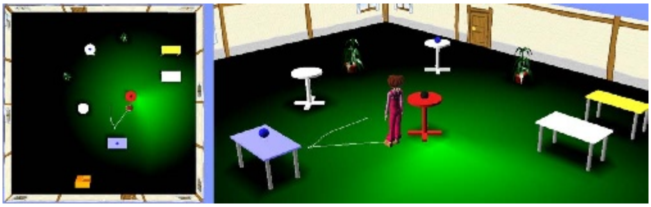


Fig. 10. "Hidari no shiroi tukue wo miro" (Look at the white table of the left.) Macro planner selected the white table from three white tables in the scene.

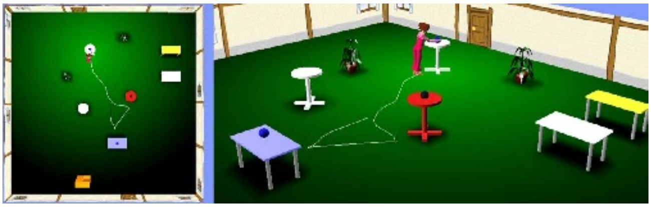


Fig. 11. "Migi no aoi tama wo kiiroi tukue ni oke" (Pick up the blue ball of the right and put it on the yellow table.) Macro planner selected the blue ball from two blue balls and micro planner generated the path in detail.Natchan walks toward the blue ball and picks it up.

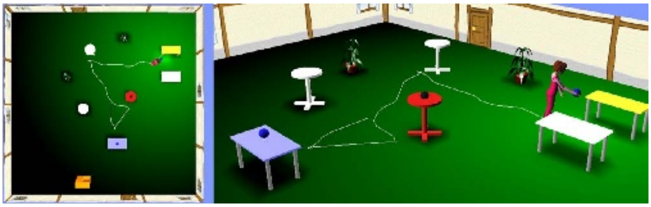


Fig. 12. "Migi no aoi tama wo kiiroi tukue ni oke" (Pick up the blue ball of the right and put it on the yellow table.) After picking the blue ball,takes it to the yellow table.

6 Conclusion

We developed a system which generates plausible animations according to natural language commands by quantifying spatial expressions and evaluating motion paths. In our system we treat spatial constraints as SPACE objects regardless of whether they are explicitly specified by the user or implied by the scene configuration. The macro planner uses SPACE objects as symbolic entities while the micro planner uses them to obtain numeric values. Using this mechanism, we can evaluate the attainment of the plan's goal and the success of the obstacle avoidance scheme collectively.

More natural motions can be performed by projecting the map of the motion database to the constraint's potential field. We are planning to improve the system which can accept more complex spacial expressions and generate more various motion according to the scene. We are planning to improve the system to accept more complex spatial expressions and generate more variable motion. For example, we want situated agents to be able to act in various styles of motion, fully exploiting their virtual space surroundings, such as passing under tables or along the crevice between walls.

References

1. Badler, N.I., Palmer, M.S., Bindinganavale, R.: Animation control for realtime visual humans. Communication of the ACM (1999) 65–73
2. Bindiganavale, R., Schuler, W., Allbeck, J., Badler, N., Joshi, A., Palmer, M.: Dynamically altering agent behaviors using natural language instructions. Autonomous Agents 2000 (2000) 293–300
3. Tanaka, H., Tokunaga, T., Shinyama, Y.: Animated agents capable of understanding natural language and performing actions. In: Life-Like Characters. Springer (2004) 429–444
4. Tokunaga, T., Koyama, T., Saito, S., Okumura, M.: Bridging the gap between language and action. In: proceedings of IVA2003, Intelligent Virtual Agents (2003) 127–135
5. Saito, S., Imoto, T., Nakajima, M.: Motiongraph: a technique of action generation for an autonomous character. The Journal of the Society for Art and Science **1** (2001) 22–29
6. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: proceedings of ACM SIGGRAPH 2002. (2002)
7. Liu, F., Liang, R.: Motion path synthesis for intelligent avatar. In: Workshop on Intelligent Virtual Agent. (2003) 141–149

Human-Scale Interaction with a Multi-projector Display and Multimodal Interfaces

Naoki Hashimoto, Jaeho Ryu, Seungzoo Jeong, and Makoto Sato

Precision and Intelligence Laboratory, Tokyo Institute of Technology,
4259 Nagatsuta, Midori-ku, Yokohama 226-8503, Japan
{naoki,jaehoryu,jeongzoo}@hi.pi.titech.ac.jp, msato@pi.titech.ac.jp
http://sklab-www.pi.titech.ac.jp/

Abstract. We propose a novel multi-projector display for immersive virtual environments. Our system named “D-vision” has a hybrid curved screen which completely covers user’s view and immerses the user in displayed virtual environments. The D-vision also has a human-scale haptic- and locomotion interface by which the users of the D-vision can walk with their own foot and interact with virtual objects through their own hands. In this paper, we show an overview of the D-vision and technologies used for the human-scale haptic- and locomotion interfaces. We also illustrate some applications using the interfaces effectively.

1 Introduction

Recently, many immersive displays have been developed for virtual reality, education, industry, entertainment, etc. These displays originated from the CAVE [1] which surrounds users with large flat screens. Although the design concept of the CAVE was quite simple, it could realize highly immersive virtual environments effectively. However, the reality of virtual environments does not only depend on visual quality. Though CAVE-like displays can present immersive virtual environments with high-quality stereoscopic images, users can not interact with the objects placed in those environments by their own hands. Haptics information, like a sense of touch, is an important factor for users to feel presence as in the environments. Bodily actions like a walking are also used to enhance the reality. In the CAVE-like displays, some controllers like a game pad or joystick are used to navigate virtual environments. These input devices are effective and easy to use, but is no intuitive. For example, actual walking actions in virtual environments give users a feeling as they are walking in the real world. A feeling of distance is also influenced by the bodily actions like a walking. In human-scale virtual environments, these feelings are important factors to produce the reality of virtual environments.

In this paper, we propose a novel multi-projector display which has haptic and locomotion interfaces for immersive virtual environments.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 23–30, 2004.

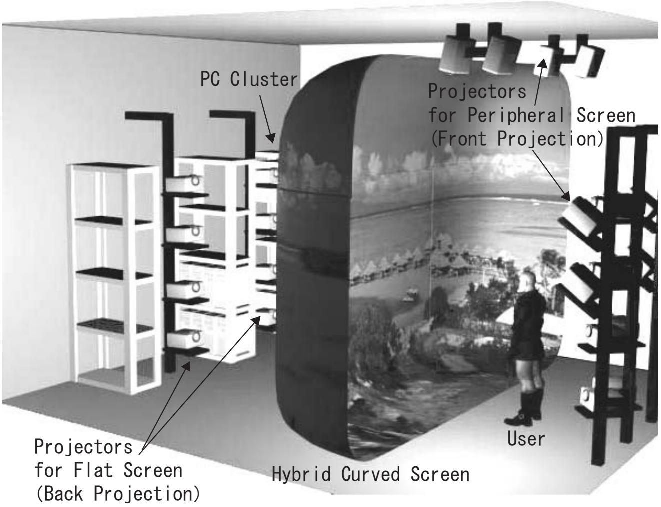

PC Cluster
Projectors for Peripheral Screen (Front Projection)
User
Projectors for Flat Screen (Back Projection)
Hybrid Curved Screen

Fig. 1. An overview of the D-vision.

2 Multi-projector Display of D-vision

This section describes the display part of the D-vision in detail. This part is designed as a visual component of the whole system that provides a virtual experience with human-scale bodily input and force feedback. In order to use limited space efficiently, our screen design attempts to realize a large screen which completely covers user's field of view, and reduce installation space. An overview of the D-vision is shown in Figure 1. Large and high resolution images are generated with a PC cluster composed of 24 commodity PCs, and projected with multi-projector strategy. A hybrid curved screen and artful projector arrangement realize reduced installation spaces. The details of these components are described in the following.

In the D-vision, a hybrid curved screen, which adopt flat screens for central view and curved screens for peripheral view, is based on the structure of human's eyes. In human's eyes, central view is used to perceive the outer world more precisely with high resolution input, and peripheral view is used to detect movements of objects in the outer world with low resolution, but wide view angle input. Therefore, a flat fresnel - lenticular stereoscopic screen is used in the central area of the hybrid screen for high quality image projection. And, in the peripheral area, simple curved screen made with fiberglass reinforced plastic (FRP) is used to realize a wide view angle. Special materials for image projection are not required because the peripheral area is not for high quality image projection. The size of the hybrid screen amounts to 6.3m (width) × 4.0m (height) × 1.5m (depth).

We use orthogonal linear polarized lights to project each image for left and the right eyes. Stereoscopic images are projected on the central flat screen and

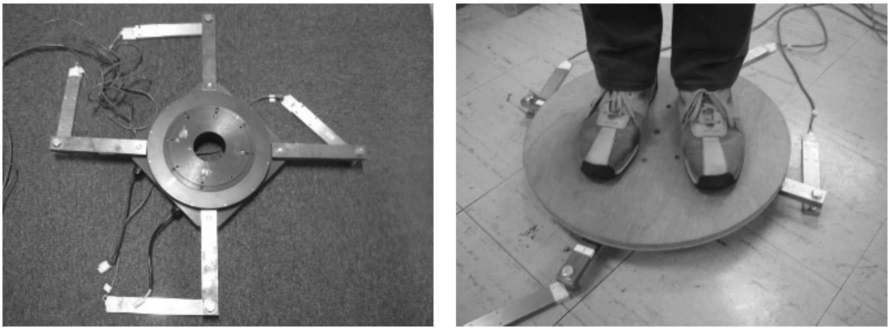


Fig. 2. A locomotion interface for the D-vision.

the upper and lower cylindrical parts of the peripheral screen. As shown in Figure 1, the central flat part of the screen is for rear projection with 8 projectors with SXGA, 1280 × 1024 pixels resolution. The remaining part of the screen is for front projection with 16 projectors with XGA, 1024 × 768 pixels resolution.

3 Locomotion Interface

The D-vision needs locomotion interfaces with step-in-place movements for two reasons. One is that bodily input enhances reality of virtual environments. The other is that view directions of users have to be controlled properly so as not to see the outside of screens because the D-vision covers only the half of the user's surrounding. We have developed a locomotion interface with a linear motor for controlling the user's view direction, and pressure sensors for detecting user's movements. The left of Figure 2 shows a structure of the interface. A center part is a linear motor which has enough torque to move users smoothly. Around the motor, 4 pressure sensors protected with iron frames are placed. In the D-vision, as shown in Figure 3, the interface is buried into the floor screen. And, on the motor, circular plate made with wood is installed to step on it. By painting the plate with the same silver paint as D-vision's peripheral screens, the interface is connected seamlessly with the D-vision. The interface is easily used by stepping on it with no wearable devices and making step-in-place movements. In immersive virtual environments, inputs without wearable devices like a joystick, wand and some sensors, are highly effective to boost the reality.

The principal function of the interface is to detect user's step-in-place movements. By using the result from the 4 sensors, a center of gravity of users is calculated in real-time. In tracks of the center of gravity, a step-in-place movement is represented as a swing movement which moves from right to left and from left to right. By comparing this simple swing movement, the step-in-place movement is easily and precisely detected only with pressure sensors. User's view direction is also detected with the supposition that the view direction is generally perpendicular to the swing direction of the center of gravity. So, turn-in-place

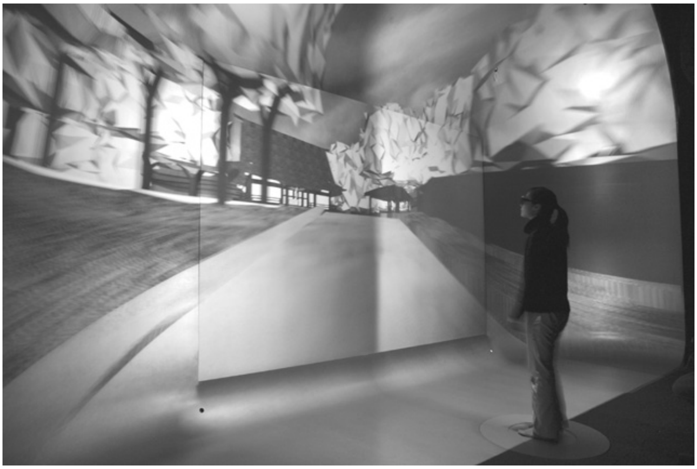


Fig. 3. A walkthrough with step-in-place movements. The scene projected on the D-vision includes steps, and the view of a user is changed according to his steps.

movements are acceptable input for the interface. A jumping and a squatting are still facile by tracing the fluctuation of the vertical load applied to the sensors.

Controlling the user's direction is also important function of the locomotion interface for the D-vision. As shown in Figure 1, the outside of the screen is viewed when users turn to the peripheral area of the hybrid screen. This situation is fatal for the D-vision because surrounding the user's view completely with images is a role of it. Therefore, the locomotion interface with a linear motor is required. In such a situation, the motor turns the users compulsorily and quietly to the central area of the hybrid screen. Physical rotation of the users and rotation of their view in virtual environments is synchronized, so they feel no incongruity of five senses. This function also contributes to the downsizing of total installation spaces for the D-vision. Without surrounding users completely with physical screens, user's view is actually surrounded with virtual environments.

An example of scenery simulation using the D-vision is illustrated in Figure 3. In this Figure, a user is going up the steps which connect to a shrine built on a hill. At first, the user can't see the shrine at all. According to his steps, the view of him is dynamically changed. And he can see the shrine gradually from the top part of it. Because the users use some control device to navigate in traditional immersive displays, they can't obtain enough feeling of being there. However, in the D-vision, our proposed locomotion interface can overcome this drawback by using step-in-place movements which require bodily inputs with comfortable fatigue. That feeling with the inteface also contributes to the sense of distance which is an important factor to apply the D-vision to architectural analysis [2]. In this way, the locomotion interface with the actual movements is essential for human-scale virtual environments.

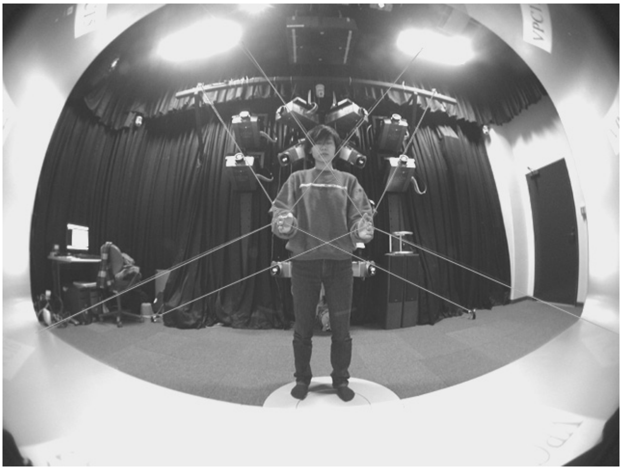


Fig. 4. A haptic interface “SPIDAR-H”. From motors to user’s hands, thin strings are strained to transfer torque generated with the motors.

4 Haptic Interface “SPIDAR-H”

For immersive virtual environments, interaction with haptic information is significant to maintain and enhance the sense of presence in those environments. In the D-vision, a haptic interface driven by some motors and strings is developed for human-scale interaction. The interface named “SPIDAR-H” is based on a haptic interface “SPIDAR” for desktop operation [3]. The structure of the SPIDAR-H is quite simple and highly scalable because it only needs some motors, which generate torque for haptics, and some strings which transmit haptics to the users. The SPIDAR-H has two rings put on the user’s finger on each hand, and each ring has four strings by which motor torque is transmitted. The motors have a rotary encoder which counts the length of the string from the ring to the motor. By using the length of the four strings, the position of the ring is calculated, and the force displayed to the users is controlled as they interact with the virtual object by their own hand directly.

An actual user wearing the SPIDAR-H is shown in Figure 4. Because SPIDAR-H is a human-scale interface, installation is a little complicated. In this system, total 8 motors for both hands are placed as surrounding the users. 4 motors placed in the front side of the users are fixed behind screens, and the strings are tensed through a small whole on the curved peripheral screen. The other motors are placed behind the users by using a frame for projectors. The strings never prevent the users from immersing into virtual worlds with surrounding images. And the flexibility of that strings enables the users to perform various motions freely.

As a typical use of the SPIDAR-H, direct interaction with our hands is realized in the D-vision, and we can operate virtual objects with the sense of touch.

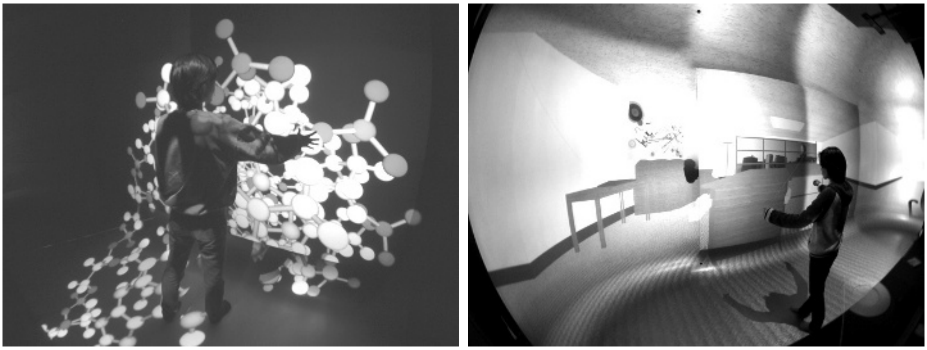


Fig. 5. Applications with the SPIDAR-H. Users can manipulate virtual objects intuitively. Haptic information intensifies immersion into virtual environments.

The left of Figure 5 shows an example of molecule visualization [4]. We can intuitively change the position of stereoscopically displayed molecules with the SPIDAR-H. The interface is also applied for education with representation of intermolecular force. The right of Figure 5 also illustrates the direct manipulation in a virtual office with the SPIDAR-H. Users can touch and move those human-scale objects with haptic information via the SPIDAR-H. A common input device like a joystick is not suitable for that kind of operation because of its limited degree of freedom. Handling human-scale objects as in the real world enhances reality of the virtual environment. This feature is quite basic and effective, but is not realized in most of conventional human-scale virtual environments.

5 Interaction with a Human-Scale Virtual Human

The D-vision is a novel immersive projection display which has multi-modal interfaces as described in above sections. Users can walk around with their own step-in-place movements, and touch and interact with virtual objects through their own hands. In order to illustrate effectiveness of these approaches, many applications are required to exploit the environments.

One of these challenges is human-scale interaction with a realistic virtual human which can behave as a real human and interact via haptic information [5] as shown in Figure 6. The implementation of the virtual human at the present time is focused to play catch with an actual human. Motion data of the virtual human is previously captured by traditional magnetic motion capture system, and stored into a SQL server. Users wearing the SPIDAR-H can interact with force feedback, and that force transmitted from the user determines virtual human's motions by selecting and modifying stored motion data in real-time. The motions of the virtual human affect virtual objects and the user oneself as force feedback. Figure 7 shows an example of this process. In Figure 7, the virtual human generates its own motions with force reaction using the SPIDAR-H. If

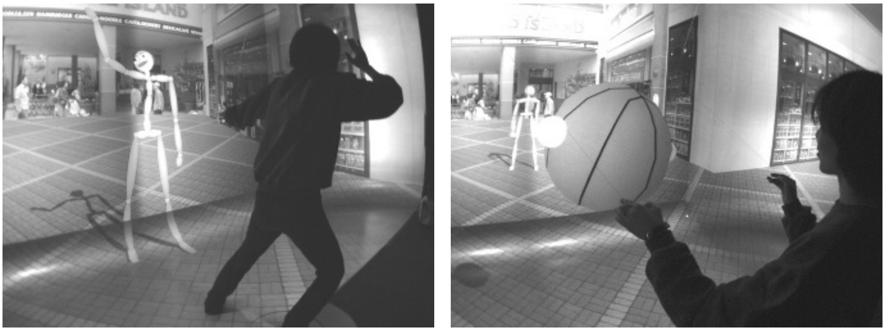
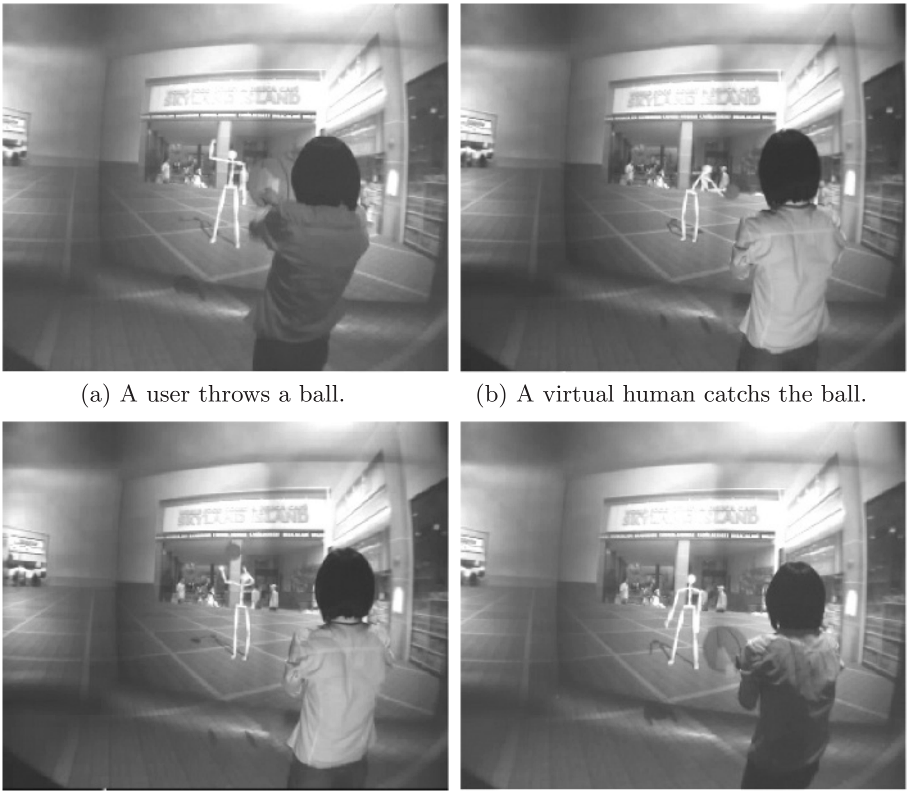


Fig. 6. Interaction with a human-scale virtual human. Users can play catch with the virtual human via the SPIDAR-H.



(a) A user throws a ball.

(b) A virtual human catchs the ball.

(c) The virtual human throws the ball.

(d) The user catchs the ball.

Fig. 7. An example of playing catch with a virtual human.

users throw a ball faster, the virtual human also perform big reaction to catch the ball. If the users throw the ball to the difficult direction to catch, the virtual human sometimes miss in catching the ball. If the users move with the locomotion interface, the virtual human reacts to the movement and changes the throwing motion. In other words, the motion of the virtual human is reactive to the users' motion. So we call it a "reactive virtual human".

Thus, the immersive virtual environment with the multimodal interfaces is suitable for next-generation interaction between human and computers. We have a plan to do a lot of trials to realize effective applications with this environment.

6 Conclusions

In this paper, we mentioned a novel immersive multi-projector display with haptic and locomotion interfaces. In this system, large images surrounding users are parallely generated with a PC cluster. A sense of the presence is enhanced with the human-scale haptic and locomotion interfaces for direct bodily input. It is illustrated that this combination innovates the reality of virtual environments compared with traditional systems biased for visual effects.

As our future plan, we will develop software which enables users to create applications easily and efficiently for the D-vision, and also try to apply this proposed system for many kinds of fields.

References

[1] Cruz-Neira, C., Sandin, D.J., DeFanti, T.A.: Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. In: Proc. SIGGRAPH '93. (1993) 135–142

[2] SOEDA, M., OHNO, R., RYU, J., HASHIMOTO, N., SATO, M.: The Effects of Architectural Treatments on Reducing Oppressed Feelings Caused by High-rise Buildings. In: Proc. 6th European Architectural Endoscopy Association. (2003) 28–35

[3] Ishii, M., Sato, M.: A 3D Spatial Interface Device Using Tensed Strings. Presence **3** (1994) 81–86

[4] Murayama, J., Bougrila, L., Luo, Y., Akahane, K., Hasegawa, S., Hirsbrunner, B., Sato, M.: SPIDAR G&G: A Two-Handed Haptic Interface for Bimanual VR Interaction. In: Proc. EuroHaptics2004. (2004) 138–146

[5] Jeong, S., Hashimoto, N., Sato, M.: A Novel Interaction System with Force Feedback between Real- and Virtual Human. Proc. International Conference on Advances in Computer Entertainment Technology (ACE2004) (2004) 61–66

Entertainment Applications of Human-Scale Virtual Reality Systems

Akihiko Shirai[1], Kiichi Kobayashi[1], Masahiro Kawakita[1], Shoichi Hasegawa[2],
Masayuki Nakajima[2], and Makoto Sato[2]

[1] NHK Engineering Services,Inc.
1-10-11, Kinuta, Setagaya-ku, Tokyo, 157-8540, Japan,
shirai@mail.com, http://adv3d.jp/
[2] Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan

Abstract. This paper describes three applications of human-scale virtual reality in demonstrative systems. The first application is a demonstrative system, the “Tangible Playroom”. It was designed as a computer entertainment system for children. It provides virtual reality entertainment with a room-scale force-feedback display, an immersive floor image, and real-time physics simulator. Children can play using their whole bodies to interact with a virtual world controlled by software using rigid body dynamics and the penalty method. The second application,“the labyrinth walker,” was designed as a virtual exploration system for children’s museum. Its step-in-place capability can provide a ‘walkable’ walk-through virtual reality environment with no worn interfaces. The third application regards photo-realistic virtual TV sets for high-definition television (HDTV) production. It can provide a real-time high-quality 3D synthesis environment using high dynamic range images (HDRI), global illumination, a HDTV depth-measuring camera called “Axi-Vision” and wire based motion control camera for real-time synthesis. In this paper, we report on these applications’ possibilities and give abstracts on their technology.

1 Computer Entertainment System Using Human-Scale VR

1.1 Tangible Playroom

The “Tangible Playroom” was designed as a future demonstrative computer entertainment system for children. By “Tangible” we mean “graspable” or “perceptible by touch”. It is an important experience for children. By “Haptics”, or “touchable virtual reality”, we mean that render stimuli touch into a virtual world. However, there are no good VR application systems for children that use haptic hardware, as far as we know. This project is thus focused on providing interesting haptic experiences for children. Figure 1 is a picture of the Tangible Playroom. The idea is that children can play with it in their rooms at home.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 31–38, 2004.

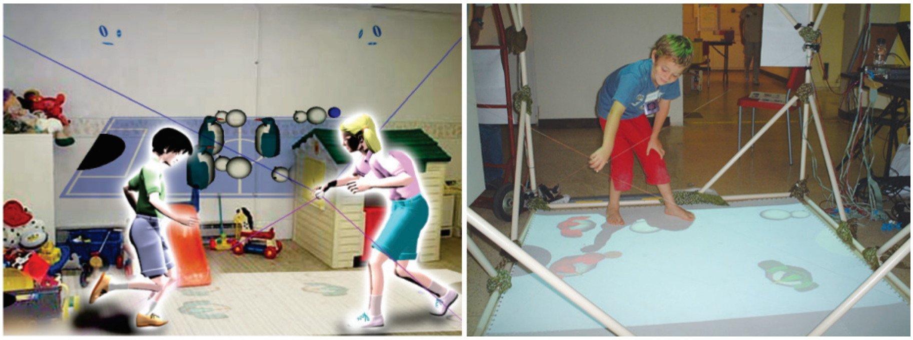


Fig. 1. "Tangible Playroom", a sketch (left) and a prototype (right)

They can interact with it by using their bodies. The game scenes are projected on the floor and walls. Its image is very large, and the children can walk directly on the screen while they play with it. The wires are links to the haptic devices internal structures. These human-scale VR systems can be turned off to discourage unlimited play.

System configuration. The system incorporates a human-scale haptic device and a large display. The large "walkable" floor screen enables the players' to move about freely. The 3D position from the tangible grip is calculated on a server PC using input from four lengths of encoder motors. The lengths are fed into the real-time rigid-body dynamics engine, which stores all the location, velocity, inertia, and behavior information for the virtual world. When it detects a collision with the floor or other characters, it uses the penalty method to generates a reaction force for force feedback via the tangible grip. All the characters in this world are driven by rigid body dynamics. Any virtual characters can move themselves autonomously based on a force vector generated by an A.I. engine. A game judge enforces rules such as scoring and time outs. Sound effects are generated according to the real-time rigid body dynamics engine based on the output of the penalty method. A multi-projection function accommodates larger displays and extra displays using networked PCs. The projected image is generated in real-time using OpenGL or DirectX. The number of projectors is variable. This software is based on a cluster real-time rendering system for CAVE-style immersive displays.

SPIDAR and haptic rendering. The Tangible Playroom is a room-scale haptic display system. To realize its force-feedback via tangible grip, the haptic system is based on "SPIDAR" (SPace Interface of Artificial Reality) [1]. SPIDAR usually uses a ring to indicate the force to users. In our system, we had to focus on the safety and convenience of children so we decided to use a cork ball as the Tangible grip.

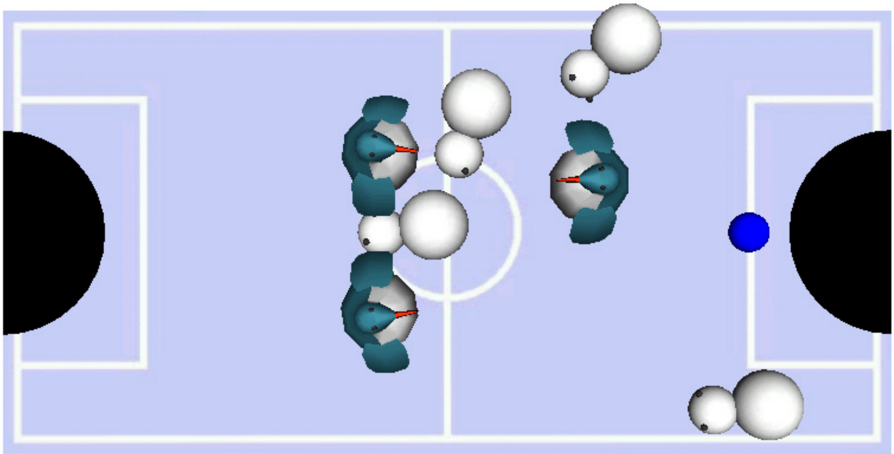


Fig. 2. “Penguin Hockey”, a demonstration content for Tangible Playroom

Demonstration Content. “Penguin Hockey” is a simple 3D hockey game content for the Tangible Playroom. It has an ice rink, four pucks, three penguins and two goals. The puck is shaped like a snowman. The children’s team (right side in Fig. 1) has one autonomous penguin, whereas their opponents, the enemy team, has two autonomos penguin players. The children thus are to help the underdog penguin in this game situation. All the objects behave according to the rigid body dynamics using the penalty method, each with a weight and a center of gravity. When a player interacts with the computer-generated characters, he or she feels the impact of the puck and the force of body checks. When penguins block the player, they check using full body movements. If a player checks one of them, they make the exclamations depending on the check’s force. The pucks and penguins have the same shape in the collision model, and the pucks have a higher center of gravity and a lighter weight compared with the penguins. This game is similar to interactive bricks, each with their own will. Playing with the penguins, passing the puck skillfully, and experiencing physical contact should be of interest to the players.

1.2 Labyrinth Walker

Locomotion interface with a floor screen. “The labyrinth walker” was designed as a virtual exploration system for children’s museums. It can provide a walk-through virtual reality environment with no interfaces that have to be worn. Photo-realistic interactive images of virtual worlds are projected onto the floor screen. Under the screen, there is an embedded locomotion interface using a linear motor-driven turntable and four pressure sensors between the turntable and the floor. When the player steps-in-place on the image, the sensors detect his

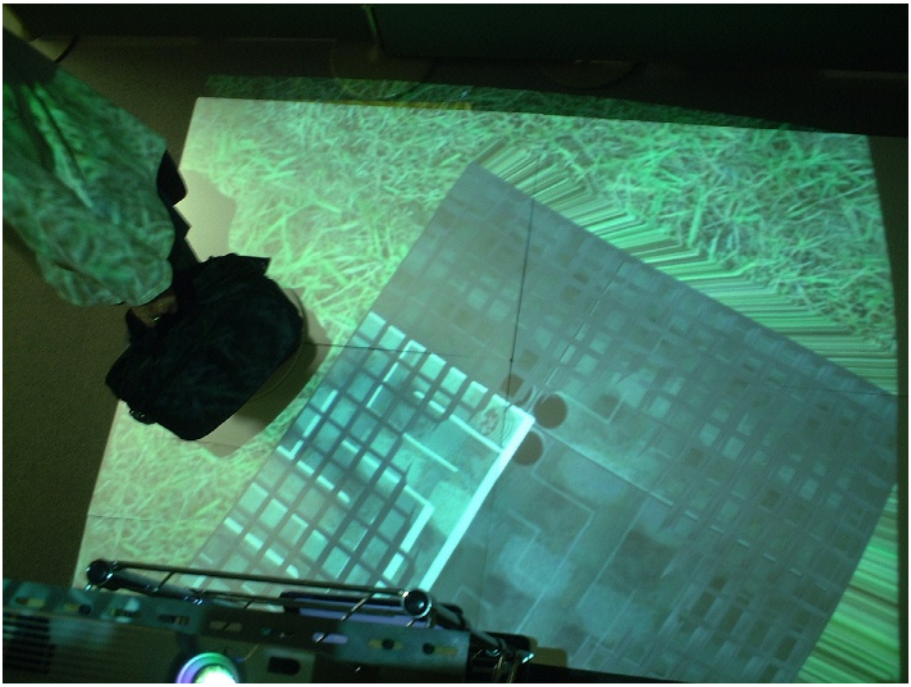


Fig. 3. The Labyrinth Walker

or her movement and orientation. The player's turning actions are then canceled by the turntable's to keep the player facing the front of screen.

The virtual scenes are written in VRML. All of behaviors involving collision and falling were developed with "Springhead", the C++ software development environment for virtual reality. The original locomotion interface system provides continuous visual feedback despite the limitations of the screen. The use of smart-turntable walking platform lets users perform life-like walking motions in a seamless manner and without wearing an interface. The interface can be easily integrated into most large-screen virtual environments. Even if the screen size is limited, the system delivers a continuous surround display. A number of children's museums have expressed their interest in purchasing this system.

2 Photo-Realistic Virtual TV Sets Systems

Photo-realistic computer graphics are difficult to achieve in a real-time graphics environment. Moreover, high-performance computers are needed to make human-scale virtual worlds sufficiently interactive. Consequently, most of VR systems are rendered by abstracted graphics images. The NHK Science and Technical Research Laboratory has studied the basic technologies for a next-generation TV production environment for making high-quality visual content.

In the stance of our research group, human-scale virtual reality systems mean to new methods of video production using real-time high-quality computer graphics with interactive techniques in TV studio sets.

2.1 High Dynamic Range Image Based Archiving and Rendering

The cinema industry uses high dynamic range images (HDRI) for image based lighting and rendering. HDRI describes a wide range of intensity by using multi-graded exposed photographs. Its images can archive the information about the light environment of TV studio sets. Figure 4 shows HDRI images of different lighting environments. We have developed a global illumination rendering

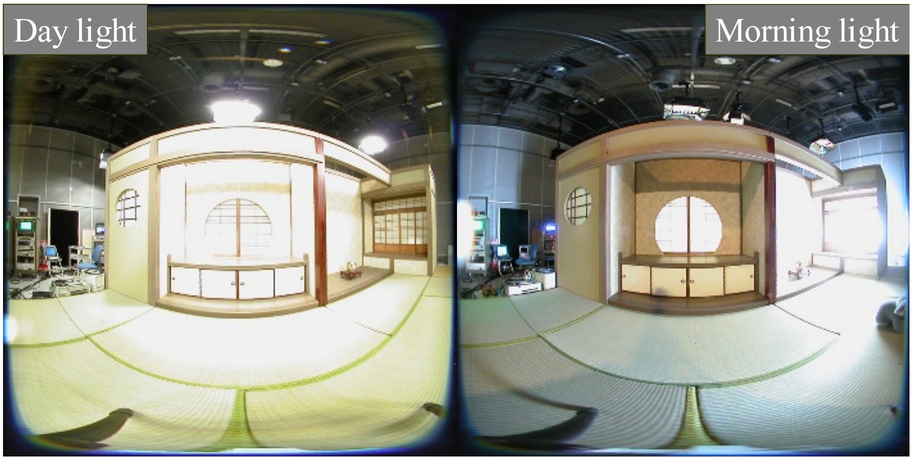


Fig. 4. High dynamic range images in different light environments

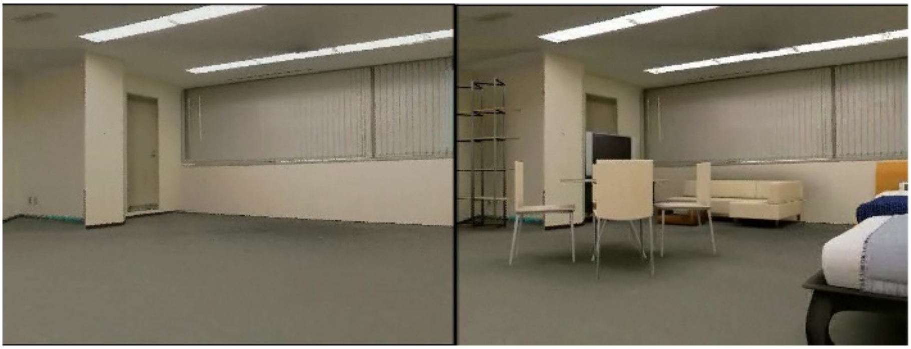


Fig. 5. Original scene (left) and artificial furnitures (right) using HDRI and global illumination rendering

system, “OptGI” for HDRI light sources. Figure 5 compares artificial images rendered with HDRI (right) and the original scene (left) by OptGI.

2.2 Virtual Shadow Casting Using Depth Camera

“Axi-Vision” is a special HDTV camera invented by Masahiro Kawakita. It can simultaneously take depth grayscale images of objects in the frame and match them to RGB pixels while operating at the full rate for HDTV movies (30 fps). This camera system has two HDTV cameras and infrared LED arrays. A dichroic prism separates these coaxial optical systems. The main system is for taking normal RGB images. The other is composed of high-definition CCD camera and a specially developed image intensifier (I.I.). The I.I. acts as an ultrahigh-speed (1 nanosecond) shuttering device with high resolution. The image including reflected light intensity by modulated LED illumination contains the depth-to-surface, orientations and reflection conditions. The ratio of the two images describes the distance from the camera to any surface in the field of view. Figure 6 contains the original background image and an artificial rabbit with her shadow. There are eleven paper plates on which the shadow should be stepped. Depth information recorded by Axi-Vision was used to make the survey model of the background.

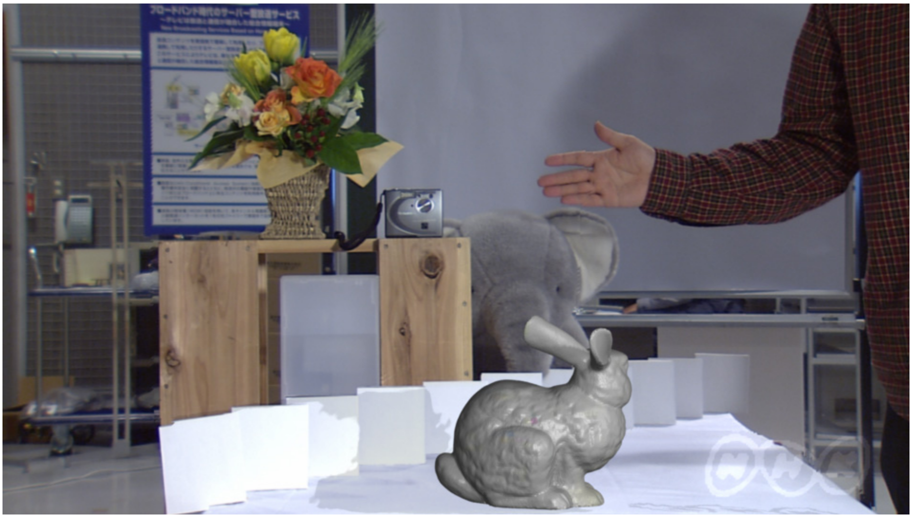


Fig. 6. Virtual shadow cast on depth image

2.3 Wire Based Motion Control Camera

Photo-realistic virtual TV sets needs real-time rendering system with least 6 DOF (degree of freedom) that contains 3 transitions (x, y, z) and 3 rotations (pitch, yaw, roll) input interface for fact camera information in three dimensional space to synthesis final images. In current technology, the rotation information can detect by tripod with mechanical encoders but transition is difficult to detect without huge mechanism such as a crane or hanger. These camera-tracking technologies are called as motion control or capture camera.

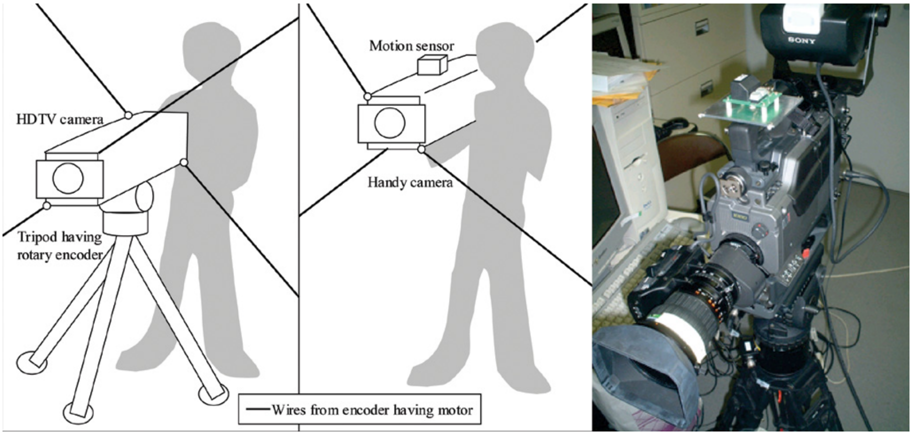


Fig. 7. Prototypes of wire based motion control camera

Figure 7 are concept pictures of wire based motion control camera. A motion sensor or mechanical encoder detect its rotation and transition tells absolute location. When the camera operator collides to an invisible virtual set in fact world, these wires tell to him/her via force feedback. Wires are not interference in TV studio rather than laboratory and lighter mechanical position detection has an advantage for using specialized camera such as Axi-Vision, infrared PSD, sensors or computer vision.

3 Conclusion

So far, our work has demon strated the practicality of human-scale VR systems in computer entertainment and TV production. This technology shouldn't be limited only CAVE style displays or visual environment. In this paper, we've just shown some suitable demonstrative application to both of fields. However each applications use some common important VR technologies such as physics engines, rendering and haptics. We expect the basic VR technologies like real-time physics engines, displays, motion tracking, haptic and photo-realistic CG rendering will be used to make new environments that are advances upon the current industrial.

Acknowledgements. This work was supported by the National Institute of Information and Communications Technology (NICT) of Japan. Global illumination renderer, "OptGI" was cooperatively developed with OptGraph, Inc.

References

1. M. Sato, Y. Hirata and H. Kawarada. SPace Interface Device for Artificial Reality-SPIDAR. The Transactions of the Institute of Electronics, Information and Communication Engineers (D-II), J74-D-II, 7, pp. 887–894, July, 1991
2. Akihiko Shirai and Makoto Sato, "Tangible Playroom: An entertainment system using a haptic interface and body interaction", 6th Virtual reality international conference, IEEE VRIC 2004 Proceedings, pp. 93–99, May, 2004
3. Makoto Sato, Laroussi Bouguila et al., "A New Step-in-Place Locomotion Interface for Virtual Environment With Large Display System", ACM SIGGRAPH 2002 Emerging Technologies, 2002
4. http://www.springhead.info/
5. S. Hasegawa, N. Fujii, Y. Koike, and M. Sato, "Real-time Rigid Body Simulation Based on Volumetric Penalty Method", Proc. of Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 326–332, 2003
6. A.Shirai, K.Kobayashi, M.Kawakita, S.Saito, M.Nakajima: A new archiving system for TV studio sets using depth camera and global illumination, NICOGRAPH International 2004, pp. 85–90, 2004
7. M. Kawakita, K. Iizuka, T. Aida, H. Kikuchi, H. Fujikake, J. Yonai, and K. Takizawa: Axi-vision camera (Real-Time Depth-Mapping Camera), Applied Optics, Vol. 39, pp. 3931–3939, 2000
8. Masahiro Kawakita, Keigo Iizuka, Tahito Aida, Taiichirou Kurita, and Hiroshi Kikuchi, Real-time three-dimensional video image composition by depth information, IEICE Electronics Express, Vol. 1, no. 9, 237–242, 2004

Analysis and Synthesis of Latin Dance Using Motion Capture Data

Noriko Nagata[1], Kazutaka Okumoto[1], Daisuke Iwai[2],
Felipe Toro[2], and Seiji Inokuchi[3]

[1] School of Science and Technology, Kwansei Gakuin University,
2-1 Gakuen, Sanda, Hyogo 669-1337, Japan
{nagata,okumotokazutaka}@ksc.kwansei.ac.jp
http://ist.ksc.kwansei.ac.jp/~nagata/

[2] Graduate School of Engineering Science, Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
{iwai@sens,toro@yachi-lab}.es.osaka-u.ac.jp

[3] Faculty of Human and Social Environment, Hiroshima International University,
555-36 Gakuendai, Kurose-cho, Kamo-Gun, Hiroshima 724-0695 Japan
inokuchi@ieee.org

Abstract. This paper presents an analysis of natural movement in Latin dance and a synthesis of dance motions making use of the outcome. The isolation movement of shoulders and hips in Latin dance was extracted quantitatively, and a dance animation with different isolation levels was synthesized, using a motion graph editor.

1 Introduction

With the rapid progress in the architectural technology of virtual environments, further research and development in technology to express the natural movement of virtual humans is being called for, from the aspects of quality and cost. In particular, in educational contents, such as navigation and sports science, and in industrial applications such as design support, and in the field of interactive media, technology which can synthesize and edit movement and technology which can form databases for motion data that already exists will be necessary.

We have extracted the characteristics inherent in various human motions and are researching animation creation technology making use of this [1,2]. This paper discusses the extraction of natural movement in Latin dance and the creation of an animation making use of the outcome. In dance, it is often said that the Japanese do not have as good rhythmical sense as foreigners. In particular, as expressed by the expression "good tempo", Latin dances are full of rhythm and motion. There are various causes for this difference such as cultural background and we cannot jump to conclusions, but there is a finding [1] that a motion unfamiliar to the Japanese people, called isolation, is involved. Through our advanced two-dimensional research, we have confirmed a phase difference in the movement of shoulders and hips which is a characteristic of the movement of Latin people. Here, in order to further conduct precision analysis, we obtained Latin dance

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 39–44, 2004.

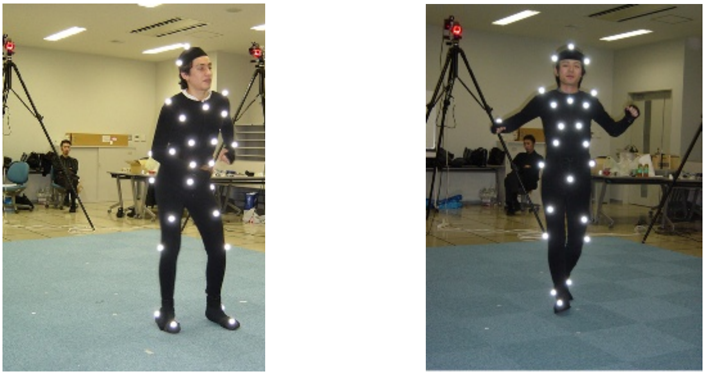


Fig. 1. Photos of subjects.

movement using motion capture and attempted to extract the difference in the characteristics of movement of people experienced in dancing (Latin people) and people inexperienced in dancing (Japanese people).

As a result, it was confirmed that isolation between the shoulder and hip takes place in a three-dimensional way in the way Latin people dance. Further, using this result, we synthesized a dance animation. To do this, we edited the bvh data of experienced people and of inexperienced people by blending them at a specified ratio using a motion graph editor [3]. In order to confirm that the created animation can express Latin movement, we made a three-dimensional display of dance animation in a human-scale virtual reality (VR) environment. The results evaluated as to whether they were like Latin movement or not.

2 Measurement and Analysis of Dance Movement by Motion Capture

We measured the dance motions of people experienced in Latin dance (Latins) and people inexperienced in dancing (Japanese), using motion capture. The motion capture system was composed of 8 digital cameras (640×480 pixels, 60Hz) and real time capture software. The subjects were asked to wear suits with 31 marks and dance to the music. The music, a Merengue piece, which is Latin, was selected (Figure 1).

As shown in Figure 2, from the dance movement obtained, the angles of rotation of shoulders and hips (the inclinations of the line connecting the joints of both shoulders (a) and both hips (b)) were calculated. This was divided into the change in the rotational component in the plane vertical to the floor and the change in the rotational component in the horizontal plane parallel to the floor.

Figure 3 shows the angles of rotation of shoulders and hips in the vertical plane. Figure 3b shows that people inexperienced in dancing almost always assume the same angle, whereas in Figure 3a, people experienced in dancing maintain a phase difference of about 90 degrees between the shoulders and hips.

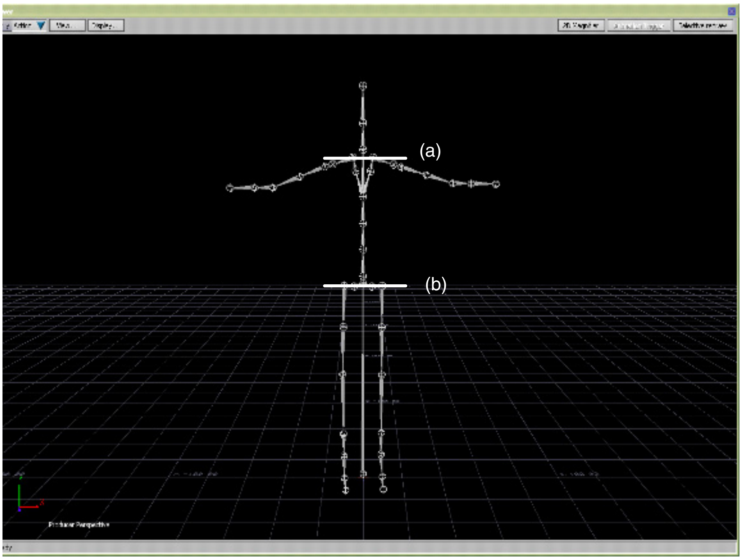


Fig. 2. The angles of rotation of shoulders and hips.

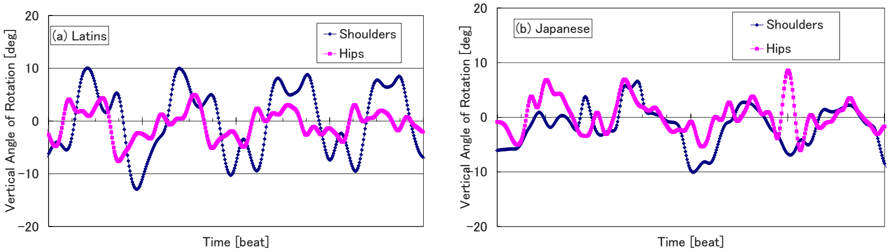


Fig. 3. The angles of rotation of shoulders and hips in the vertical plane.

Figure 4 shows the angles of rotation of shoulders and hips in the horizontal plane. Similarly Figure 4a indicates that the movement of Latins keeps a phase difference of about 30 degrees between the shoulders and hips, no phase difference was observed in the Japanese dance.

These can be interpreted as an example of isolation between shoulders and hips moving independently, and it is considered that one of the characteristics of Latin dance has been successfully extracted.

Figure 5 shows the center of gravity of hips in the vertical plane. The up and down movements of the Latins are very small compared with the Japanese. This can also be explained as isolation, one of the characteristics of Latin dance.

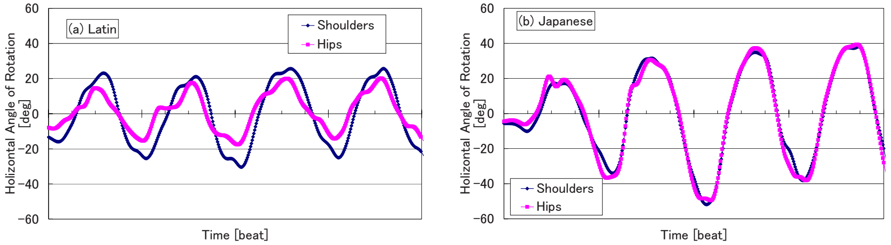


Fig. 4. The angles of rotation of shoulders and hips in the Horizontal plane.

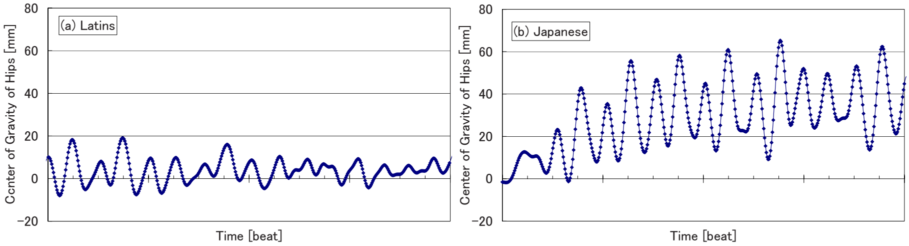


Fig. 5. The center of gravity of hips in the vertical plane.

3 Synthesis of Dance Animation

In order to confirm that isolation between shoulders and hips is one of the characteristics of Latin dance, we created a dance animation using our motion data and evaluated it.

We used a motion graph editor to synthesize the dance motion. The motion graph editor can connect the motion data expressed in a multiple bvh form in serial/parallel. The joint can be linearly interpolated by line blending in any ratio using the motion blend function. For instance by using a pre-measured walking movement and a running movement, it is possible to form an animation that can express the natural shift from walking to running.

Using this blending function, motion data with a difference in the isolation level between the shoulders and hips was synthesized. To be specific, the dance motions from the same cycle of the dance of an experienced person and an inexperienced person were cut out and created by blending them in a certain ratio. Then three motions, motion A (blend ratio 9:1), motion B (5:5, as shown

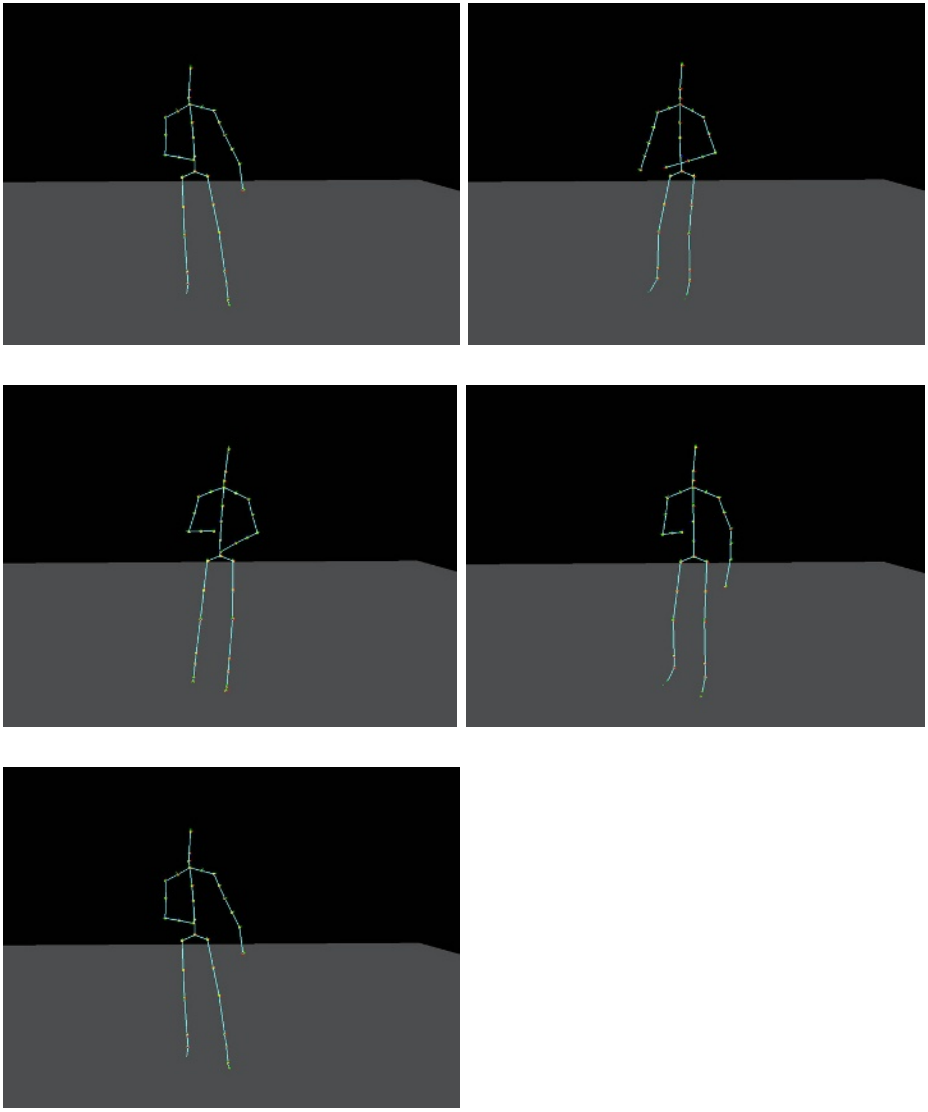


Fig. 6. An example of Synthesized Latin dance motion.

in Figure 6), motion C (1:9) were randomly presented to subjects who were asked to put the Latin characteristics in order. For this presentation, motion data converted from bvh type to xml type was displayed in three-dimensions, using VR software (Omegaspace). The result of the ranking by 3 subjects was from top down, motion A, motion B and motion C. Thus it was confirmed that the dance motion expresses Latin characteristics.

4 Conclusion

The isolation movement of shoulders and hips in Latin dance was extracted quantitatively. In order to verify whether this characteristic resembled Latin movement, a dance animation with different isolation levels was synthesized, using a motion graph editor. As a result of the evaluation, it was confirmed that Latin movement can be expressed appropriately. In future, we plan to analyze synchronism with music and study getting into the rhythm.

References

[1] Iwai, D., Toro, F., Inokuchi, S.: Analysis of Dance Motion: Japanese and Latins. In: Proc. SI2002. (2002) 299–300

[2] Kamisato, S., Yamada, K., Tamaki, S.: Sensitivity evaluation and three-dimensional motion analysis of arm motion in dance. (2004) 1–4

[3] Saito, S., Imoto, T., Nakajima, M.: Motiongraph: a Technique of Action Generation for an Autonomous Character. (2002) 22–29

Web-Based Telepresence System Using Omni-directional Video Streams

Kazumasa Yamazawa[1], Tomoya Ishikawa[1], Tomokazu Sato[1], Sei Ikeda[1],
Yutaka Nakamura[2], Kazutoshi Fujikawa[2], Hideki Sunahara[2], and
Naokazu Yokoya[1]

[1] Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara, Japan,
{tomoya-i,yamazawa,tomoka-s,sei-i,yokoya}@is.naist.jp,
http://yokoya.naist.jp/

[2] Information Technology Center, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara, Japan,
{yutaka-n,fujikawa,suna}@itc.naist.jp,
http://inet-lab.aist-nara.ac.jp/index_eng.html

Abstract. Recently, the telepresence which allows us to experience a remote site through a virtualized real world has been investigated. We have already proposed telepresence systems using omni-directional images which enable the user to look around a virtualized dynamic real scene. The system gives the feeling of high presence to the user. However, conventional systems are not networked but are implemented as stand-alone systems that use immersive displays. In this paper we propose a networked telepresence system which enables the user to see a virtualized real world easily in network environments. This paper describes the system implementation using real-time and stored omni-directional video streams as well as its experiments.

1 Introduction

Recently, there are many researches of telepresence that acquires a dynamic real world into a virtual world and enables the user to be immersed in the remote environment [1]. The telepresence can be applied to various fields such as entertainment, medical service, and education.

We have already proposed telepresence systems which use omni-directional camera and enable the user to look around the scene in order to increase the presence in telepresence [2,3]. The systems acquire and transfer the remote omni-directional scene by an omni-directional camera, and show the user view-dependent images in real time. They have no significant delay from the rotation of user's head to the presentation of images. They also enable the user to look around the remote scene widely. However, the conventional systems are implemented as stand-alone systems with immersive displays and require special programs and equipments. It is difficult for multiple users at remote sites to look around the same scene with the conventional systems.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 45–52, 2004.

In this paper, we propose a new telepresence system which enables multiple users on network to look around remote environments captured by omnidirectional cameras. The system uses web-browser and enables users to see omnidirectional videos interactively.

Section 2 describes the proposed system. Section 3 describes experiments of the system with live video and stored video. Finally, Section 4 summarizes the present work.

2 Web-Based Telepresence System Using Omni-directional Videos

The schema of proposed system is illustrated in Fig. 1. Omni-directional videos are stored in a remote server and are acquired by the viewer which is started by a web browser. The user looks around the omni-directional video contents on the web browser.

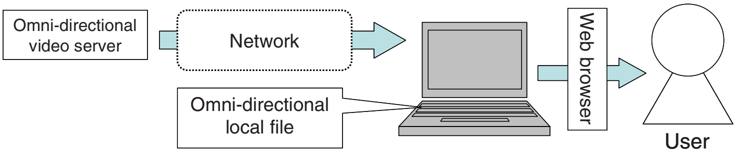


Fig. 1. Overview of web-based telepresence system.

2.1 Omni-directional Video Viewer on Web Browser

A web browser is one of the most popular network applications. Especially, Internet Explorer installed on Windows machines can execute various application programs by a JAVA applet or Active-X for providing users with interactive contents on a web page. Moreover, the JAVA applet and the Active-X programs can be easily distributed by an automatic install function. Thus we implement an omni-directional video viewer for telepresence on a web browser in this study.

The omni-directional video viewer which shows the user omni-directional video contents needs the functions of GPU (graphic processor unit) and is implemented by Active-X. The omni-directional video viewer is started by the web browser, converts the omni-directional video to the common perspective video, and shows the user the common perspective video on the web page. The system uses a hardware texture mapping function for converting video in real-time by the method [2].

When the user access the web page which provides an omni-directional video con-tent, the omni-directional video viewer implemented by Active-X is installed automatically. The user can see an omni-directional video content without care of an omni-directional camera type, parameters of camera, file-path of the content, and so on, because a content provider embeds them in a HTML file.

The user can look around the omni-directional video by using a mouse-drag operation. The omni-directional video viewer is installed only by opening the web page, and the omni-directional video can be seen easily. In the other implementation, the user can look around the omni-directional video through a HMD (Head Mounted Display) with a gyro sensor.

2.2 Omni-directional Video Contents

There are two kinds of omni-directional video contents in present implementation: stored video contents encoded in advance and live video contents encoded in real-time. The stored video contents mainly consist of high-resolution omni-directional videos heavy for network. Note that stored video contents can be provided as an on-demand-service. The live video contents are used for the purpose of providing multiple users with the same contents simultaneously just like TV broadcasting. It is difficult to transfer a high-resolution video because of the limit of standard network bandwidth. The user can see the live video contents acquired by an omni-directional camera and transferred immediately. It can be transferred to many sites by multi-cast without increasing the network load.

3 Experiments

We have implemented the proposed system and experiments with both stored video contents and live video contents. Stored video contents consist of high-resolution videos obtained by using an omni-directional multi-camera system. Live video con-tents are acquired by using an omni-directional camera mounted on a car and that are transferred via wireless and wired network in real time.

A) Stored Video Contents

We acquired the omni-directional video by an omni-directional multi-camera system; Ladybug (see Fig. 2) [3] and stored it in a PC for presentation (see Table 1). The camera unit of Ladybug consists of six cameras (Fig. 2(left)): Five configured in a horizontal ring and one pointing vertically. Fig. 2(right) shows a storage unit which consists of an array of HDD. The camera system can acquire video covering 75% of the full spherical view. The acquired video has the size of 3840 × 1920 pixels and is captured at 15fps. In this experiment, we shrink the video to 1024 × 512 pixels because of the limit of HDD-access-speed of the PC. We used MPEG-1 video and MPEG-1 layer 2 sounds for the formats of the video.

Fig. 3 shows a window shot of the web browser. The user can look around the scene on the web browser. The PC can playback the video at 30fps. The user can pause and fast-forward the video with the stored video contents.

We have also implemented a view-dependent presentation system (see Fig. 4) with a HMD and a gyro sensor for realizing more rich presence. The gyro sensor is Inter-Trax2 made by INTERSENSE. It can acquire the user view-direction at 256Hz. The user can look around the omni-directional scene without significant delay.

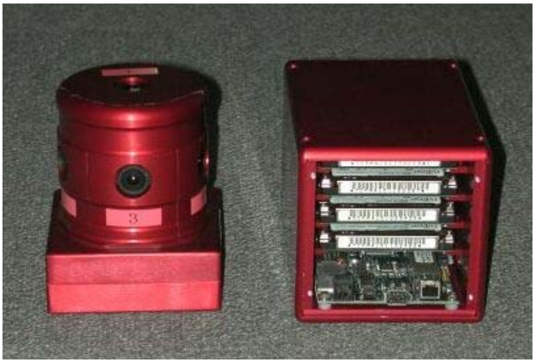


Fig. 2. Omni-directional multi-camera system; Ladybug.

Table 1. PC for presentation of stored video contents.

CPU	Pentium4 2GHz
Memory	512MB
Graphics card	ATI RADEON9700pro
OS	Windows XP

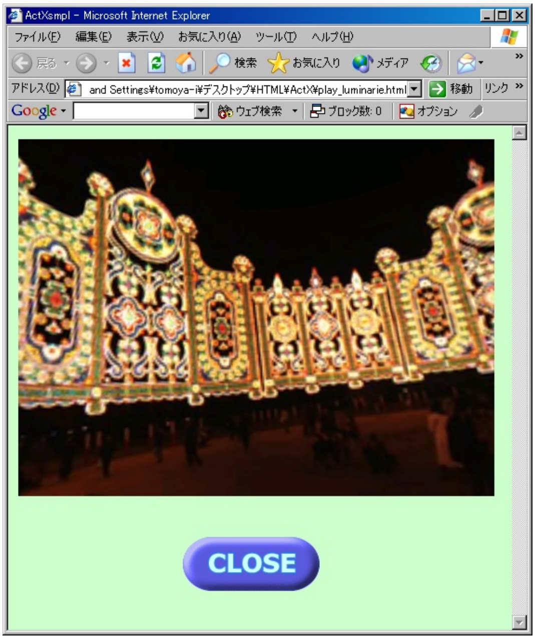

ActXsmpl - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り メディア
アドレス(D) and Settings¥tomoya-i¥デスクトップ¥HTML¥ActX¥play_luminarie.html 移動 リンク
Google ウェブ検索 ブロック数: 0 オプション
CLOSE

Fig. 3. Window shot of omni-directional video viewer with stored video content.

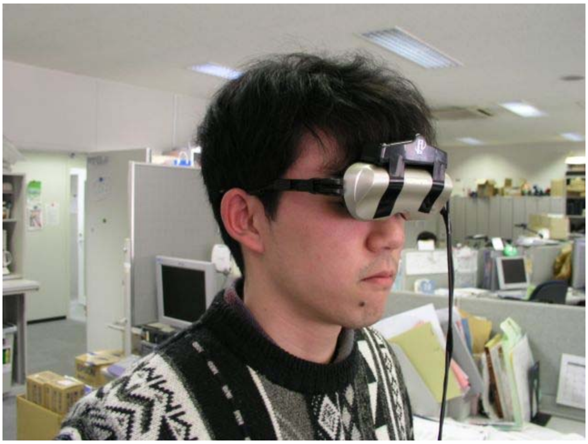


Fig. 4. HMD (Head Mounted Display) with gyro sensor.

B) Live Video Contents

In the experiment of live video contents, the system consists of a car which mounts omni-directional camera HyperOmni Vision [4], multicast relay server of video, omni-directional video viewer, and network (see Fig. 5).

The car mounting the omni-directional camera acquires omni-directional progressive video surrounding the car, running in our campus. The acquired omni-directional video is transferred to a PC for encoding in the car through iLink. The PC encodes the omni-directional video (640 × 480 pixels, 30fps) to Windows Media Format (1Mbps) by Windows Media Encoder [4]. The encoded omni-directional video is transferred to the indoor relay server through IEEE802.11a or g network via wireless network. Table 2 describes the configuration of the system for acquiring omni-directional video streams. Fig. 6 and Fig. 7 show the car which mounts the omni-directional camera and the system which mounted on the car.

The transferred omni-directional video is received by the multicast relay server of omni-directional video. The relay server distributes the omni-directional video by multicast such as RTSP protocol. The distributed omni-directional video is seen by using the same omni-directional video viewers as for stored video contents on the web browsers. When many viewers receive the omni-directional video, the load of network does not increase because of using not unicast but multicast.

In the experiment, actually four PC received the distributed omni-directional video. The four users could look around the scene in arbitrary directions. Fig. 8 shows examples of windows shots of the omni-directional video viewer. The video is dis-played on the web browser at 30fps. The time delay between the acquisition and the presentation omni-directional video is 10 seconds. In this time, the both transmitting and receiving network loads of the relay server are 1Mbps. When the number of received omni-directional video viewer increased, the network load did not increase.

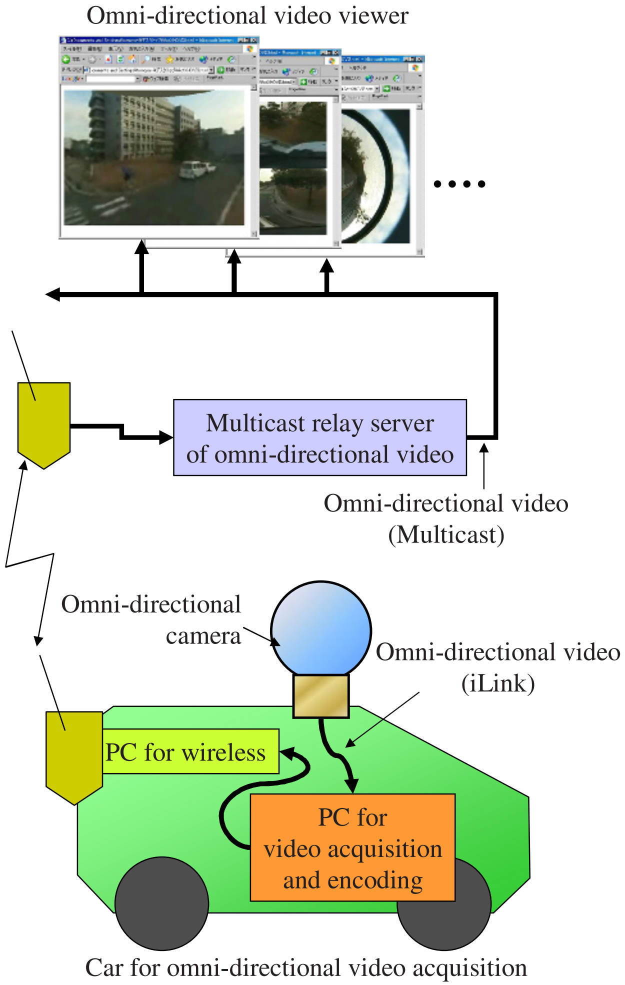

Omni-directional video viewer
....
Multicast relay server of omni-directional video
Omni-directional video (Multicast)
Omni-directional camera
Omni-directional video (iLink)
PC for wireless
PC for video acquisition and encoding
Car for omni-directional video acquisition

Fig. 5. Telepresence system with omni-directional camera mounted on car.

Table 2. Facilities for omni-directional video acquisition in outdoor environments.

Omni-directional camera	SONY DCR-TRV900 + Hyperboloidal mirror (field of view : 30 degrees upper)
PC for video acquisition and encoding	Pentium4 2.53GHz Memory 1GB WindowsXP
Wireless network	IEEE802.11a and g
Car	Nissan ELGRAND (see Fig.6, Fig.7)

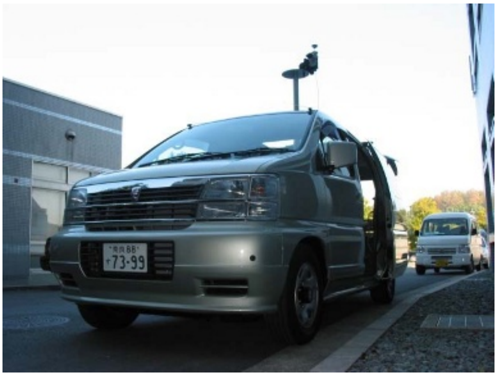


Fig. 6. Appearance of car for omni-directional video acquisition.

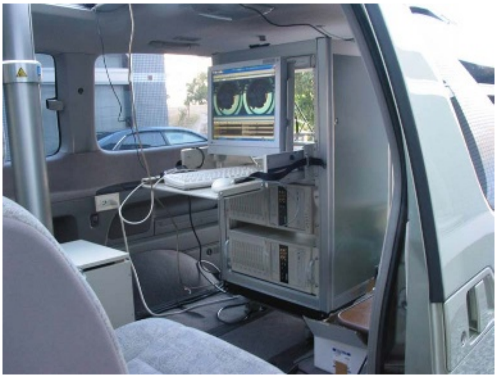
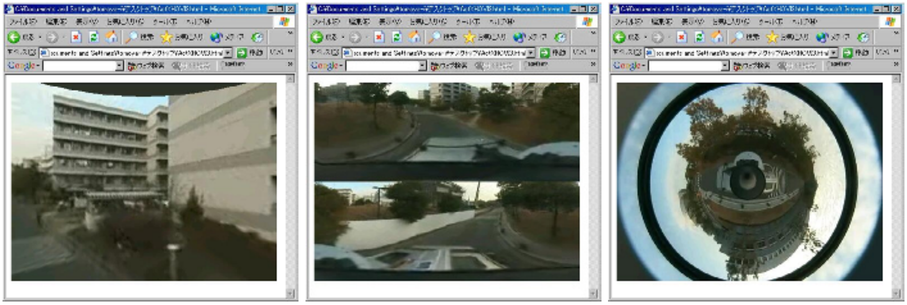


Fig. 7. Window shots of omni-directional video viewer.



Common perspective image

Panorama image
(Cylindrical image)

Omni-directional image
(Same as input image)

Fig. 8. System of car for omni-directional video acquisition.

4 Conclusion

We have developed a new networked telepresence system which easily enables multiple users to look around a remote scene with omni-directional camera. The system uses a web-browser, and enables users to see omni-directional video such as common video. In the experiment of stored video contents, the user could see the high-resolution omni-directional video. In the experiment of live video contents, the omni-directional video was distributed through wireless and wired network by multicast protocol, and multiple users could look around the scene in arbitrary directions in real time.

In the experiment of live video contents, the omni-directional camera was NTSC. The resolution of the omni-directional video was low. Thus the presence was not rich enough. On the other hand, in the experiment of stored video contents, an omni-directional multi-camera system was employed for acquiring a high-resolution video. It was not suitable for distributing an omni-directional live video. Therefore omni-directional camera should be high resolution and should not be multi-camera system. The omni-directional HD camera can acquire omni-directional high-resolution live video.

In the experiment of live video contents, the delay between the acquisition and the presentation omni-directional video is 10 seconds. It is difficult to use the system for communication with a remote user. In future work, we should reduce the delay in transmitting an omni-directional video stream.

References

1. Moezzi, S., Ed.: Special issue on immersive telepresence. IEEE MultiMedia. **4** 1 (1997) 17–56
2. Onoe, Y., Yamazawa, K., Takemura, H., and Yokoya, N.: Telepresence by Real-time View-dependent Image Generation from Omnidirectional Video Streams. Computer Vision and Image Understanding. **71** 2 (1998) 154–165
3. Ikeda, S., Sato, T., Kanbara, M., and Yokoya, N.: Immersive telepresence system using high-resolution omnidirectional movies and a locomotion Interface. Proc. SPIE Electronic Imaging. **5291** (2004)
4. Yamazawa, K., Yagi, Y., and Yachida, M.: Omnidirectional imaging with hyperboloidal projection. Proc. Int. Conf. on Intelligent Robots and Systems. **2** (1993) 1029–1034
5. Microsoft Corporation, Windows Media Encoder 9 Series. `http://www.microsoft.com/windows/windowsmedia/9series/encoder/`

Wide View Surveillance System with Multiple Smart Image Sensors and Mirrors

Ryusuke Kawahara, Satoshi Shimizu, and Takayuki Hamamoto

Tokyo University of Science, Department of Electrical Engineering
Shinjyuku, Tokyo 162-8601, Japan
{kawahara,satoshi,hamamoto}@isl.ee.kagu.tus.ac.jp

Abstract. We describe a wide view imaging system for surveillance which uses multiple smart image sensors and mirrors. In this system, each image obtained by the multiple sensors has no-overlapped area and is equivalent to the partial image of the wide view obtained by only imaginary sensor. Therefore depth estimation from sensor to each object is not required for combination. In this paper, we describe the wide surveillance imaging system by using random access image sensors we have designed and a FPGA. We can control the system to show panoramic image or partial image in real time. The new image sensor has useful functions for wide view imaging which are random accessing and interpolation of pixel values on quarter pitch. We show some results obtained by the chip.

1 Introduction

We have been investigating random access image sensors which are applicable to smart imaging systems [1,2]. Only the pixel values selected by control signals can be output from the random access image sensor, therefore the sensor is very effective to the imaging system by using multiple sensors.

In this paper, we describe a wide view imaging system for surveillance application by multiple image sensors and mirrors [3,4]. In this system, a panoramic image obtained by a virtual sensor is divided into multiple images obtained by corresponding sensors so that the further calculation is not required to combine the multiple images except image projection.

We propose new random access image sensor for the wide view imaging that has 128 × 128 pixels. The imaging system using the new sensors can capture and display panoramic view image or arbitrary view image and control its view angle in real time.

2 Wide View Imaging System by Using Multiple Sensors and Mirrors

2.1 Capturing Wide View and Image Synthesis

We use multiple sensors and corresponding mirrors for wide view imaging. Fig. 1 shows the cross section view in the case of eight sensors and an octagonal mirror. Although the sensors a-h are put on a plane and capturing upper views

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 53–60, 2004.

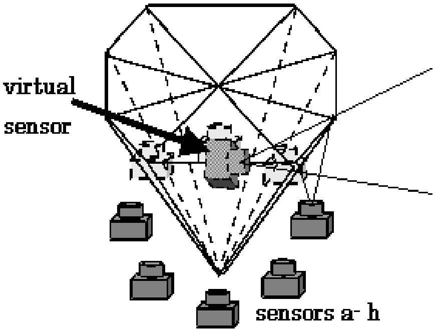


Fig. 1. Wide view imaging with eight sensors and an octagonal mirror

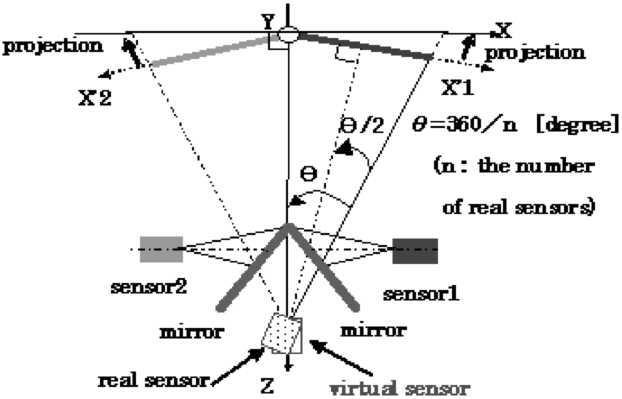
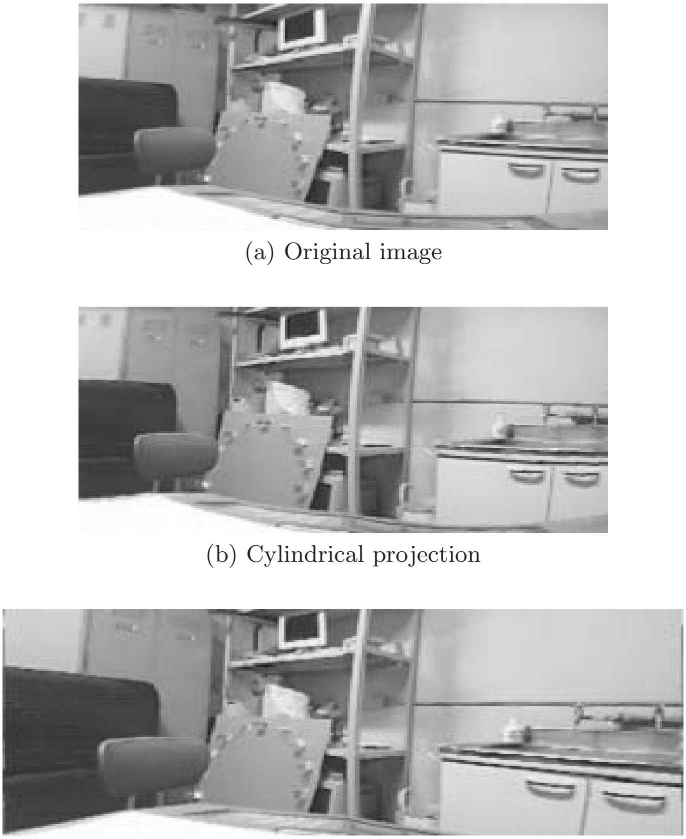


Fig. 2. Projection for combining two images

independently, the combination of eight images corresponds to the image of a virtual sensor by using the octagonal mirror. Therefore the virtual sensor has eight times as wide as the view angle of the single sensor. The combination of eight images corresponds to the wide view captured from one position and has no-overlapped area, therefore the estimation of depth from the sensor to the objects and its compensation are not necessary for combining eight images.

Although the depth estimation is not necessary, the processing of projection is required for combining various images. Fig. 2 shows the cross section view with two sensors and mirrors. Two images on X'1 plane or X'2 plane are combined and transformed to an image on X plane. For synthesis of the all-directions view image, each image is transformed cylindrically. Fig. 3 shows the simulation results of the projections. Fig. 3(a) shows the original image which consists of two adjacent images without the projection. Figs. 3(b) and (c) show the transformed images with cylindrical and plane projections. Although the processing of the projection is required for making wide view image, the positions of output pixels from each sensor are fixed for all frames.



(a) Original image

(b) Cylindrical projection

(c) Plane projection

Fig. 3. Images transformed by cylindrical projection and plane projection

2.2 Imaging System Using Random Access Image Sensor

Fig. 4 shows our prototype system for wide-view imaging. The system consists of eight normal CCD sensors positioned on a plane and an octagonal mirror. Fig. 5 shows an example panoramic image obtained by the prototype system. Fig. 6 shows the system architecture for imaging and displaying panoramic view image or arbitrary view image in real time. In the case of normal CCD sensors, image data for all pixels are memorized and the partial pixels are selected by projection map for the display. Therefore quite big circuits are necessary if the number of sensors is large. On the other hand, in the case of random access sensors, selection of the displayed pixels can be done on each sensor and only the pixel data shown on the monitor are output. According to the projection method, FPGA controls the positions of the output pixels by chip selection signals and address signals. The output image data can be directly displayed without further processing. The proposed system using the random access sensors is much small compared to the system using CCD sensors.

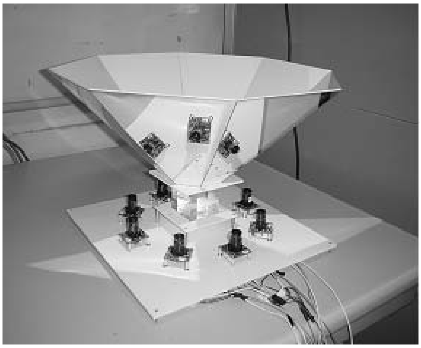


Fig. 4. Prototype system for wide view imaging with eight CCD sensors

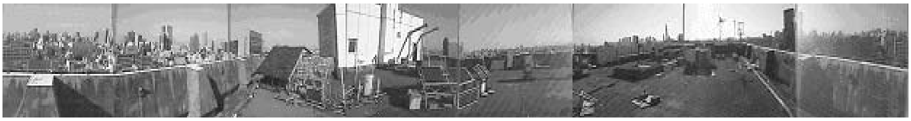
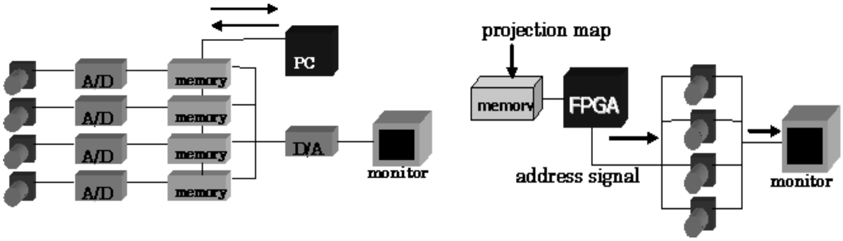


Fig. 5. Example panoramic image with eight CCD sensors



(a) The system using normal CCD sensors

(b) The system using smart sensors

Fig. 6. Comparison between normal CCD sensors and smart sensors

3 Random Access Image Sensor

3.1 Design and Implementation of the Random Access Image Sensor

Fig. 7 shows the block diagram of the random access image sensor for wide view imaging system we have designed. It consists of photodiode array, shift registers, address decoders, APS control circuits, an average circuit and an output control circuit. Pixel cell has a sample and hold circuit to integrate the pixel value at the same timing for all pixels. In the average circuit, we use 16 capacitors for interpolation of the adjacent four pixels. In the output control circuit, the pixel values are adjusted to reduce the variation between the eight sensors.

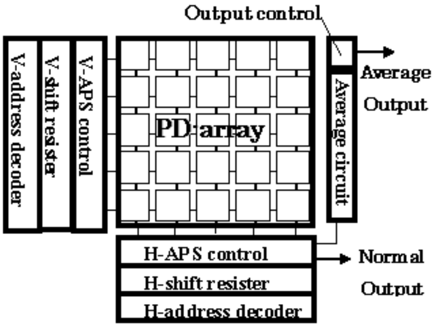


Fig. 7. Block diagram of the random access image sensor

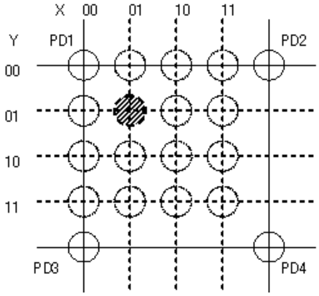


Fig. 8. Estimation of pixel value on quarter pitch

The proposed chip has the following four functions.

[1] Pixel data is output by normal shift register or address decoder for random access.
[2] Pixel value on every quarter position is estimated by averaging circuit.
[3] Each chip can be set its fixed number for sensor identification.
[4] The pixel value is amplified or offset for reduction of variation between sensors.

Fig. 8 shows the example of quarter pitch interpolation. Suppose the pixel value on (01,01) is estimated, the value PD_{ave} is calculated by next equation.

$$PD_{ave} = \frac{9 \times PD_1 + 3 \times PD_2 + 3 \times PD_3 + 1 \times PD_4}{16} \tag{1}$$

We have fabricated the chip using AMS 2-poly 3-metal 0.6 μm CMOS process. The outline of the proposed chip is shown in Table 1.

3.2 Experiments

Fig. 9 shows the example images obtained by the prototype chip. Fig. 9(a) shows a normal output image and Fig. 9(b) shows the projected output images.

Table 1. Outline of the random access image sensor

Number of pixel [pixels]	128 × 128
CMOS process [μm]	0.6
Die size [mm^2]	4.27 × 4.26
Pixel size [μm^2]	22.0 × 22.0
Fill factor [%]	22.1
Number of transistors of transducer [trs./pixel]	6
Number of transistors of interpolation [trs./chip]	537
Power supply [V]	5.0

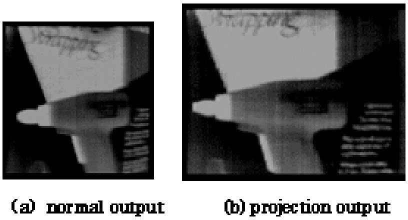

(a) normal output (b) projection output

Fig. 9. Example images obtained by the chip

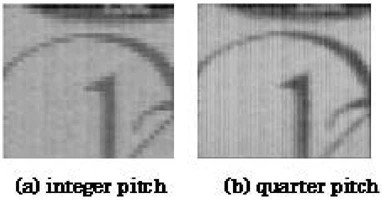

(a) integer pitch (b) quarter pitch

Fig. 10. Expanded images by average circuit

It appears that slope characters in normal output image are transformed to the straight characters in Fig. 9(b). Fig. 10 shows the expanded images in integer pitch and in quarter pitch. It appears that the slope line in Fig. 10(b) has smoother shape compared with Fig. 10(a).

4 Wide View Surveillance System

We propose a wide view surveillance system using new smart sensors as shown in Fig. 11. The system consists of an octagonal mirror, FPGA, memory and eight sensors. Fig. 12 shows the block diagram of the system. In this system, FPGA selects output pixels from the image of the eight sensors using 4bit “sensor select data” and 18bit “pixel address data”. The memory keeps the transformed data for the projection and various calibrations which are calculated by previous experiments. By using this memory, “pixel address data” including integer and decimal address X, Y is transformed into the projected and calibrated address data with no-calculation in real time. Fig. 13 shows the output images from the sensor 1 to 8.

Now we are implementing the functions of detecting and tracking of moving object on the surveillance system. In this paper, we show simulation results by

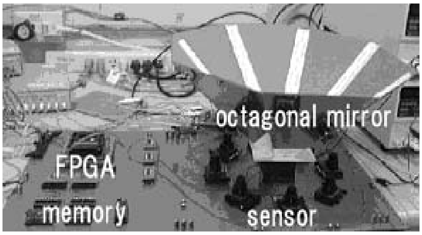


Fig. 11. Wide view surveillance system

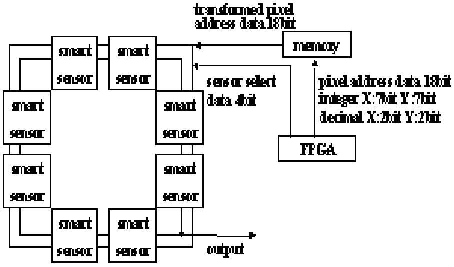


Fig. 12. Block diagram of the surveillance system

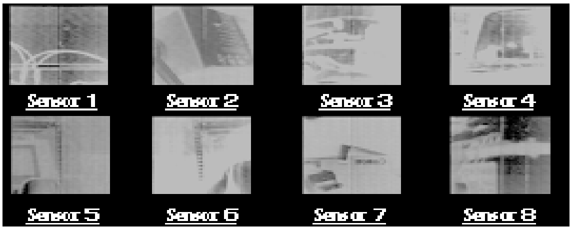
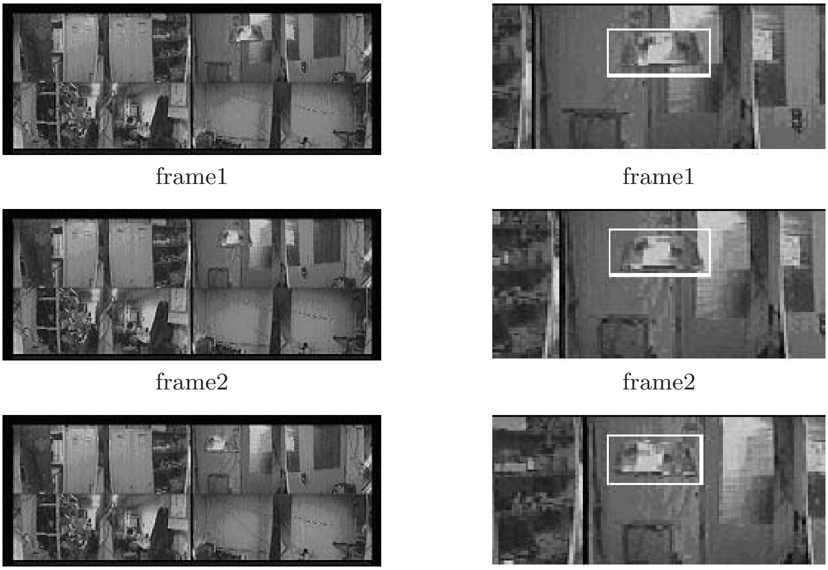


Fig. 13. Output images of sensor 1-8



frame1

frame1

frame2

frame2

frame3

frame3

(a) Panorama view in low resolution; the upper-right image is connected to the bottom-left image, the bottom-right image is connected to the upper-left image

(b) Detection of moving object and zooming image in high resolution

Fig. 14. Example results of surveillance application

using the prototype system. The surveillance system selects the arbitrary pixel data from eight sensors pixel by pixel, therefore various parameters such as view angle, spatial resolution and temporal resolution can be controlled freely. Fig. 14 shows the example results of objects tracking system. If there is no moving object, panorama view images as shown in Fig. 14(a) are output in low resolution to control the amount of output data. On the other hand, if moving object is detected, the partial images including the moving object are also output in high resolution as shown in Fig. 14(b). However the resolution of the panorama view images are reduced when the partial images are shown.

5 Conclusion

We have fabricated a random access image sensor which has 128 × 128 pixels for wide view imaging. We propose the wide view surveillance system with eight smart sensors and an octagonal mirror. The system can capture all-directions view or partial view in real time. We will show the results of the object tracking by using the constructing surveillance system in the conference.

References

1. R.Ooi, T.Hamamoto, T.Naemura, K.Aizawa, “Pixel independent random access image sensor for real time image-based rendering system”, IEEE Int. conf. on Image Processing (ICIP’01), TA6, pp. 193–196, 2001
2. R.Kawahara, S.Shimizu, T.Hamamoto, “Wide view imaging system by using random access”, Int’l Conference on Multi sensor Fusion and Integration for Intelligent Systems, pp. 185–190
3. Kawanishi, Yamazawa, Iwasa, Takemura and Yokoya, “Generation of High-resolution Stereo Panoramic Image by Omnidirectional Imaging Sensor Using Hexagonal Pyramidal Mirrors”, IEICE Technical Report, PRMU 97–118, 1997 (in Japanese)
4. Kenji Tanaka, Kenji Suzuki, Yasunari Suzuki, Mitsuo Isogai, Yoshiki Arakawa, Hideshi Tanaka, and Masahito Sato, “8-Miliion Pixels Ultra High Definition Images System”, IEICE Technical Report, EID 2001–36, pp. 7–12, 2001

Object Tracking and Identification in Video Streams with Snakes and Points

Bruno Lameyre and Valerie Gouet

CEDRIC/CNAM - 292, rue Saint-Martin - F75141 Paris Cedex 03
bruno.lameyre@free.fr, valerie.gouet@cnam.fr

Abstract. This paper presents a generic approach for object tracking and identification in video sequences, called SAP. The object is described with two image primitives: first, its content is described with *Points of interest* that are automatically extracted and characterized according to an appearance-based model. Second, the object's envelope is described with a *Snake*. The originality of SAP consists in a complementary use of these primitives: the snake allows to reduce the points extraction to a limited area, and the point description is efficiently exploited during the snake tracking. Such a characterization is robust to wide occlusions and can be use for object identification and localization purposes. SAP has been implemented with the aim of achieving near real-time performance.

1 Introduction

In a variety of applications of image technology, such as medical image analysis, video surveillance or scene monitoring, it is desirable to track objects in video sequences. Considerable work has been done during the past few years in object tracking. There is no theory for the segmentation of moving objects in videos, the methods depend upon the target application. When a model of the object does not exist, the encountered approaches usually focus either on image spatial structures, or on temporal tracking with trajectory estimation, or on both. Different kinds of approaches exist and are usually based on region segmentation [9], blobs [12], histograms [16], optical flow [2], points [20] or snakes [3], etc.

The paper is organized as follows: Section 2 describes our approach of object's content description which is based on points of interest. In Section 3, we remind of snakes principles before presenting a novel approach of object tracking combining snakes and such points. Experiments on video streams are presented in Section 4 to highlight the contributions of the SAP approach. Finally, we propose a natural extension to object identification and localization.

2 Object Tracking with Points of Interest

Points of interest are involved in many applications, like stereovision, image retrieval or scene monitoring. They usually represent sites where the information is considered as perceptually relevant. Many extractors have been proposed, see

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 61–68, 2004.

for example the comparison study [19]. The most popular one is probably the Harris and Stephens detector [7] and its adaptations [18,15,14].

Temporal approaches of feature point tracking exist for *point trajectory estimation.* Classically, the encountered techniques involve a cost function defined for three consecutive frames. Different linking strategies are applied to find the correspondences and optimize the trajectories. The first solution is the one of Sethi and Jain [20] called Greedy Exchange algorithm. Some improvements of this approach have been proposed [17,21]. In [4], the algorithm "IPAN tracker" described is based on the idea of competing trajectories. The paper also presents a performance evaluation of feature points tracking approaches.

Most of the approaches listed above estimate a trajectory according to a local model of trajectory. They do not exploit the visual appearance of the points to track. Since they involve a model of trajectory, they are not robust to wide deformation of non-rigid object and to wide occlusions. In this paper, we focus on spatial appearance-based tracking approaches. Such techniques do not impose any constraint on the trajectory of the point and may allow wide occlusions, as it will be demonstrated. Traditional approaches involving a *spatial description of points* come from stereovision or more recently from image retrieval applications. From the works of Koenderink [10] and Florack [5] on the properties of local derivatives, a lot of work has been done on differential descriptors. A recent performance evaluation of local descriptors [13] has shown that the descriptor SIFT proposed by Lowe [11] for object recognition performs best.

2.1 Our Approach of Point of Interest Tracking

We did not make the choice of employing the SIFT descriptor in our prototype, first because it involves a high dimensional features set (128 items for each keypoint), making it not applicable for real-time video tracking purposes. Second, this descriptor is invariant to several image transformations, making it efficient for object recognition but not optimal for video streams where consecutive frames differ by small transformations. Therefore, the characterization employed here is the local jet of the signal which is invariant to image translation. Up to order n, it can be expressed for the point (x, y) as follows:

$$J(x, y, \sigma) = \{I_{i_1 \ldots i_k}(x, y, \sigma)/k = 0, ..n\} \tag{1}$$

where $I_{i_1 \ldots i_k}(x, y, \sigma)$ represents the k^{th} image derivative relative to the $i_1 \ldots i_k$ variables (x and y) and σ the size of the Gaussian smoothing applied during the derivatives computation. Under the gaussian assumption, the similarity measure traditionally combined with these features is the Mahalanobis distance δ^2. In the rest of the paper, such a point characterization space will be noted (V_d, δ^2).

Point matching algorithm. A specific model of trajectory is not exploited here. We only suppose that the point p_i^j characterizing the object O_i of frame F_i has its corresponding point in frame F_{i+1} inside an area which is simply modelled by a circular window W_t of size t centered on p_i^j. The matching algorithm consists in finding in (V_d, δ^2) the nearest neighbor p_{i+1}^k of p_i^j, with p_{i+1}^k in $W_t(p_i^j)$. A match having a distance δ^2 higher than a given threshold is eliminated. Under some hypothesis, the threshold can be automatically chosen from the χ^2 table. Then a classical cross-matching algorithm is applied in (V_d, δ^2) in order to build a set of matches $\{(p_i^j, p_{i+1}^k)\}$ with points involved in each match at best one time. Our algorithm privileges the visual similarity of interest points. The t parameter can be viewed as a function of the velocity of the point to track. It can be estimated from the couple (p_{i-1}^l, p_i^j) of matched points, as in the approaches of point trajectory estimation. In that case, the matches also involve points which are constrained by a particular velocity from frames F_{i-1} to F_{i+1}.

3 Robust Object Tracking with Snakes

The Snakes theory was born in 1987 with the work of Kass et al. [8]. A state of art about snakes can be found in [1]. They are widely used for segmentation, shape modelling and motion tracking. A snake can be represented as a parametric curve. From a given starting position, the snake deforms itself in order to stick to the nearest salient contour. The snake behavior and its evolution are governed by a weighted combination of *internals* and *externals* forces and is computed as an energy function to minimize. Such a minimization is not easy. A numerical solution consists in considering a discrete representation of the curve and in developing an iterative algorithm. The implementation of the snake we have chosen for our prototype is classical. For the regularization of the curve, three forces are applied on each node. The first one is a *stretching* force which guarantees a certain distance between two consecutive nodes. The second one is a *bending* force which constrains the curvature on each node. The third one is the *external* force which is directly linked to the image contours. Some temporal forces can be added to help during the tracking [3]. In order to reduce computation time, we choose a determinist algorithm which reduces the total energy of the snake by reducing the energy of each node separately. This process is iteratively repeated as the snake energy decreases.

3.1 Exploiting Points of Interest to Enhance Snake Tracking

The view O_i of an object in a frame F_i can be described with a set of interest points noted P_i and with a discrete snake noted S_i. The complete object characterization obtained is the couple (P_i, S_i). In this section, we propose a method consisting in exploiting the P_i features to make the snake tracking more robust.

Let consider two sets of points P_i and P_j characterizing two views O_i and O_j of the same object in two frames F_i and F_j. Matching these two sets (or subsets) allows to estimate an image transformation $T_{i,j}$ existing between O_i and O_j. $T_{i,j}$ can be used at two different levels of the snake tracking: first between two consecutive frames F_i and F_{i+1}. The snake S_{i+1} can be initialized with $T_{i,i+1}(S_i)$, before optimizing it for the view O_{i+1}. Second, $T_{i,j}$ can be exploited to make the snake tracking much more robust against wide occlusions: according to the

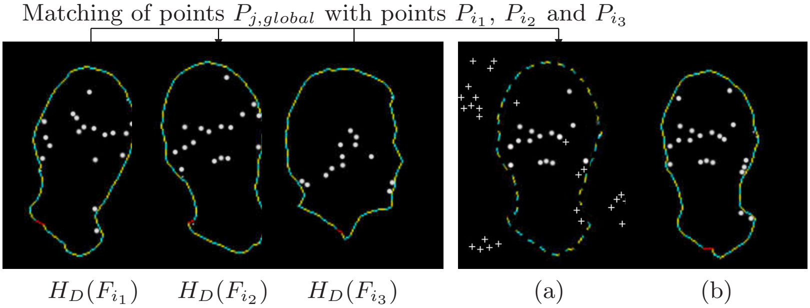

Matching of points $P_{j,global}$ with points P_{i_1}, P_{i_2} and P_{i_3}
$H_D(F_{i_1})$
$H_D(F_{i_2})$
$H_D(F_{i_3})$
(a)
(b)

Fig. 1. Tracking of a face after a full occlusion. On the left, 3 items of the history list. On the right, (a) shows the points $P_{j,global}$ extracted on a whole frame F_j after the occlusion. • points are the P_j which better match with points of H_D (here with P_{i_2}) and + points are unmatched points. The dotted snake drawn is $T_{i_2,j}(S_{i_2}) = S_{j,init}$. (b) shows the P_j points plus the optimized snake S_j obtained from $S_{j,init}$.

object characterization we have adopted, we consider that an object becomes occulted in a frame F_i when few points P_i can be matched with P_{i-1}. In such a case, points are extracted in the whole frames $F_{j,j\geqslant i}$ as long as the object is occulted. The corresponding sets obtained are called $P_{j,global}$.

Now, let suppose that we have at our disposal the description noted $(P_{i_{ref}}, S_{i_{ref}})$ of one of the views $O_{i_{ref}}$ before the occlusion of the object, and the set $P_{j,global}$ extracted from the frame F_j when it reappears. $P_{i_{ref}}$ and $P_{j,global}$ can be compared according to an approach similar as the one detailed in Section 2.1. Here the points of the two sets are to be compared. It is not possible to directly use the point characterization based on the local jet (equation (1)) since $T_{i_{ref},j}$ can be more complicated than a small inter-frames motion. Therefore, the point characterization we consider is a combination of the local jet computed during the tracking, in order to achieve invariance to other images transformations. For example, it can be the Hilbert's differential features which are invariant to image rotation. Such a derived characterization is necessary to compare points under the hypothesis of different viewpoints between $F_{i_{ref}}$ and F_j, but makes the point characterization less selective. It is possible to enrich it by adding geometric and semi-local constraints on points as the ones proposed in [6]. In the rest of the paper, we will note $(V'_{d'}, \delta^2)$ such a feature space following from (V_d, δ^2).

It is then reasonable to suppose that the points of $P_{j,global}$ which are involved in the matches obtained give a characterization P_j of the view O_j. Then estimating $T_{i_{ref},j}$ from some of the points $(P_{i_{ref}}, P_j)$ in correspondence allows to initialize in F_j a snake with $T_{i_{ref},j}(S_{i_{ref}})$. This technique supposes that a view $O_{i_{ref}}$ which is quite similar to O_j exists and that $(P_{i_{ref}}, S_{i_{ref}})$ is available. To do that, our approach consists in storing during the tracking sub-samples $(P_i, S_i)_{i=k_1,..,k_D}$ of the object characterization in a FIFO list called H_D. Under this hypothesis, $(P_{i_{ref}}, S_{i_{ref}})$ can be chosen within H_D as the description which fits better a subset of $P_{j,global}$, according to a score SCR_H which is inversely proportional to the distances obtained between matched points. The algorithm is illustrated in Figure 1 and completely described in the next section.

Algorithm 1: Object tracking with snakes and points of interest.

```
// Initializations
- Manual surrounding of the object to track in $F_1$. It gives $S_{1,init}$;
- Optimization of $S_{1,init}$ for the object $O_1$. A refined snake $S_1$ is obtained;
- Extraction of a set of points $P_1$ in $F_1$ inside the area defined by $S_1$;
For each frame $F_{j,j>1}$ of the sequence do
    If $(P_{j-1}, S_{j-1}) \neq (\varnothing, \varnothing)$ then
        // The object was globally visible in frame $F_{j-1}$
        - Extraction of a set of points $P_j$ in $F_j$ inside an area $\mathcal{A}(S_{j-1})$;
        - Point matching of the $P_{j-1}$ set with the $P_j$ one in $(V_d, \delta^2)$;
        - $P_{i_{ref}} \leftarrow P_{j-1}$;
    else
        // The object was widely occulted in frame $F_{j-1}$
        - Extraction of a set of points $P_{j,global}$ in the whole frame $F_j$;
        - Search in $H_D$ of the $P_i$ set associated with the best score
        $SCR_H(P_i, P_{j,global})$. It involves a subset $P_j \subset P_{j,global}$;
        - $P_{i_{ref}} \leftarrow P_i$;
    end if
    If enough $P_{i_{ref}}$ points are matched with the $P_j$ ones then
        // The object is globally visible in frame $F_j$
        - Estimation of $T_{i_{ref},j}$ from the matches between $P_{i_{ref}}$ and $P_j$;
        - $S_{j,init} \leftarrow T_{i_{ref},j}(S_{i_{ref}})$;
        - Optimization of $S_{j,init}$ for $O_j$. A refined snake $S_j$ is obtained;
        - $H_D \leftarrow H_D + (P_j, S_j)$;
    else
        // The object is widely occulted in frame $F_j$
        $P_j \leftarrow \varnothing$; $S_j \leftarrow \varnothing$;
    end if
    $j \leftarrow j + 1$;
end for
```

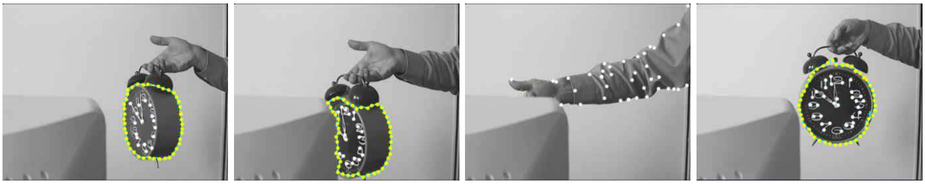
3.2 The Complete Algorithm of Tracking with Snakes and Points

The SAP algorithm proceeds as described in Algorithm 1. Points are extracted using the Precise Harris detector and characterized with the local jet in (V_d, δ^2) and the derived feature space $(V'_{d'}, \delta^2)$. They are matched according to the approach of Section 2.1. In this algorithm, the window $W_{S_{i-1}}$ considered for the points of interest extraction in frame F_i is based on the snake computed in frame F_{i-1}. The area defined by S_{i-1} only gives in F_i a first approximation of the area where to extract the points that will characterize the object. Since the points to track may have moved between the two frames, it is necessary to consider an enlarged surface. We consider a simple dilatation of the surface defined by S_{i-1}, noted $\mathcal{A}(S_{i-1})$. The size of the dilatation can be viewed as a function of the points velocity, as for the parameter t of the window W_t used during the point matching process between two frames (see Section 2.1).

4 Results of Object Tracking and Identification

The video resolution is QCIF (352 × 288) and image is acquired in YUV12 format (4:2:0). Only the Y part, containing the grey level information is exploited.

Object tracking during a full occlusion. Here, the object (a clock) completely disappears behind an obstacle. When disappeared, its trajectory does not follow the same one as before the occlusion, making a model of trajectory unusable. Figure 2 presents particular frames before, during and after the occlusion. The SAP characterizations associated are superimposed on the frames.

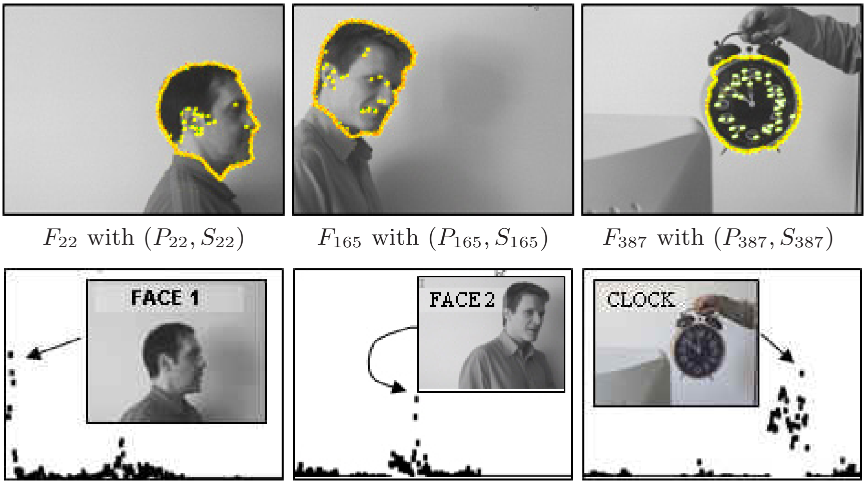


F_{378} with (P_{378}, S_{378}) F_{606} with (P_{606}, S_{606}) F_{693} with $P_{693,global}$ F_{761} with (P_{761}, S_{761})

Fig. 2. Evolution of the object characterization (P_i, S_i) during the tracking, in the presence of a full occlusion. Frames (F_{378},F_{606}), F_{693} and F_{761} have been respectively taken before, during and just after the occlusion.

Extension to object identification and localization. We propose a simple and natural extension of the SAP approach to automatic object identification and localization. Let us consider sequences where several objects O_k are tracked. If the history list H_D^k corresponding to each O_k is stored and labelled in a global H set, then H represents a file of objects labelled and characterized by their content that can be viewed as a thesaurus. Given a new sequence, the approach consists in extracting in the whole frames F_j a set of points $P_{j,global}$ and in finding the ones that fit better the ones stored in H, by using the Algorithm 1. The *identification* simply consists in giving the label(s) associated with the item $(P_i, S_i) \in H$ having the best score $SCR_H(P_i, P_{j,global})$. The *localization* consists in estimating in F_j the snake $S_{j,init} = T_{i,j}(S_i)$ and then in optimizing it to obtain S_j (let note that S_j can be considered as the starting point of a new tracking phase for the identified object). This scenario is illustrated in Figure 3 with a video containing three objects (two faces and a clock) to identify and localize from a given thesaurus of several objects.

About computation time. All the algorithms developed have been chosen to be real-time compatible. At present, the optimization phase have not yet been made but we think that real-time is achievable. The following estimations give an idea of the actual performances, based on an Intel Centrino 1.6 Ghz CPU computer:



F_{22} with (P_{22}, S_{22}) F_{165} with (P_{165}, S_{165}) F_{387} with (P_{387}, S_{387})

$SCR_H(P_i, P_{22,global})$ $SCR_H(P_i, P_{165,global})$ $SCR_H(P_i, P_{387,global})$

Fig. 3. Identification and localization of 3 objects. First line shows frames F_j of the tested video, with $j = 22, 165, 387$. Second line presents the graph of the scores SCR_H obtained between all the items P_i of H_D and $P_{j,global}$. The points presented on the images of first line are the $P_j \subset P_{j,global}$ that matched. The object identification step is represented on second line by labelled thumbnails that correspond to the items of H_D having the best scores. The localization step is done with the estimation of the snake S_j (drawn on the images of first line), that has been obtained from $T_{i,j}(S_i)$ and optimized.

- Snake used alone (as object tracker): 25 ms/frame (40 fps);
- Snake used in cooperation with feature points tracker: 80 ms/frame (12 fps);
- Time to retrieve the best candidate in H_{256}: 200 ms, depending on the number of points inserted in each history item.

5 Conclusions and Future Work

In this paper, we have presented a novel approach for object tracking in video sequences. The object to track is described by considering two generic image primitives: points of interest and snakes. No model of object nor trajectory is used to achieve the tracking. We focused our work on two particular aspects: first, we tried to develop an appearance-based point characterization the most robust possible to the variability that an image coming from a video may contain. Second, we exploited such a characterization to make the snake tracking more robust. The experiments realized on wide occlusions clearly show the relevance of the spatial description of the points we propose, when a temporal one would be lacking. In addition, we have proposed an extension of SAP to object identification and localization, which is another application that gives promising results. We are now investigating techniques for optimizing the access to the thesaurus and making it more representative.

References

1. A. Blake and M. Isard. *Active Contours*. Springer, 1998
2. G. Castellano, J. Boyce, and M. Sandler. Regularized cdwt optical flow applied to moving-target detection in IR imagery. *Machine Vision and Applications*, 11(6):277–288, 2000
3. C. Chesnaud, P. Réfrégier, and V. Boulet. Statistical region snake-based segmentation adapted to different physical noise models. IEEE *PAMI*, 21(11):1145–1157, 1999
4. D. Chetverikov and J. Verestóy. Tracking feature points: A new algorithm. In *In Proc. International Conf. on Pattern Recognition*, pages 1436–1438, 1998
5. L.M.J. Florack, B.M ter Haar Romeny, J.J. Koenderink, and M.A. Viergever. General intensity transformations and differential invariants. *Journal of Mathematical Imaging and Vision*, 4(2):171–187, 1994
6. V. Gouet and N. Boujemaa. Object-based queries using color points of interest. In *IEEE Workshop CBAIVL*, pages 30–36, Kauai, Hawaii, USA, 2001
7. C. Harris and M. Stephens. A combined corner and edge detector. *Proceedings of the* 4^{th} *Alvey Vision Conference*, pages 147–151, 1988
8. M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contours models. *International Journal of Computer Vision*, pages 321–331, 1988
9. M. Kim, J.G. Jeon, J.S. Kwak, M.H. Lee, and C. Ahn. Moving object segmentation in video sequences by user interaction and automatic object tracking. *IVC*, 19(5):245–260, April 2001
10. J.J. Koenderink and A.J. Van Doorn. Representation of local geometry int the visual system. *Biological Cybernetics*, 55:367–375, 1987
11. David G. Lowe. Distinctive image features from scale-invariant keypoints. *Accepted for publication int the International Journal of Computer Vision*, 2004
12. R. Megret and J.M. Jolion. Tracking scale-space blobs for video description. IEEE *Multimedia*, 9(2):34–43, 2002
13. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Intl. Computer Vision and Pattern Recognition*, 2003
14. Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *ICCV'01*, Vancouver, Canada, July 2001
15. P. Montesinos, V. Gouet, and R. Deriche. Differential Invariants for Color Images. In *ICPR'98*, Brisbane, Australia, 1998
16. P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Eur. Conf. on Computer Vision, LNCS 2350*, pages 661–675, Copenhaguen, Denmark, June 2002
17. V. Salari and I.K. Sethi. Feature point correspondence in the presence of occlusion. IEEE *PAMI*, 12(1):56–73, January 1990
18. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *PAMI*, 19(5):530–534, May 1997
19. C. Schmid, R. Mohr, and Ch. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000
20. I. K. Sethi and R. Jain. Finding trajectories of feature points in a monocular image sequence. IEEE *PAMI*, 9:56–73, 1987
21. C.J. Veenman, E.A. Hendriks, and M.J.T. Reinders. A fast and robust point tracking algorithm. In *International Conference in Images Processing*, 1998

Optical Flow-Based Tracking of Deformable Objects Using a Non-prior Training Active Feature Model[⋆]

Sangjin Kim[1], Jinyoung Kang[1], Jeongho Shin[1], Seongwon Lee[1], Joonki Paik[1], Sangkyu Kang[2], Besma Abidi[2], and Mongi Abidi[2]

[1] Image Processing and Intelligent Systems Laboratory,
Department of Image Engineering,
Graduate School of Advanced Imaging Science, Multimedia, and Film,
Chung-Ang University, 221 Huksuk-Dong, Tongjak-Ku, Seoul 156-756, Korea,
paikj@cau.ac.kr, http://ipis.cau.ac.kr

[2] Imaging, Robotics, and Intelligent Systems Laboratory,
Department of Electrical and Computer Engineering,
The University of Tennessee, Knoxville, TN 37996-2100, USA
http://imaging.utk.edu

Abstract. This paper presents a feature point tracking algorithm using optical flow under the non-prior training active feature model (NPT-AFM) framework. The proposed algorithm mainly focuses on analysis of deformable objects, and provides real-time, robust tracking. The proposed object tracking procedure can be divided into two steps: (i) optical flow-based tracking of feature points and (ii) NPT-AFM for robust tracking. In order to handle occlusion problems in object tracking, feature points inside an object are estimated instead of its shape boundary of the conventional active contour model (ACM) or active shape model (ASM), and are updated as an element of the training set for the AFM. The proposed NPT-AFM framework enables the tracking of occluded objects in complicated background. Experimental results show that the proposed NPT-AFM-based algorithm can track deformable objects in real-time.

1 Introduction

The problem of deformable object tracking by analyzing motion and shape in two-dimensional (2D) video is of increasing importance in a wide range of application areas including computer vision, video surveillance, motion analysis and extraction for computer animation, human-computer interface (HCI), and object-based video compression [1,2,3,4].

[⋆] This work was supported by Korean Ministry of Science and Technology under the National Research Lab. Project, by Korean Ministry of Education under Brain Korea 21 Project, by the University Research Program in Robotics under grant DOE-R01-1344148, by the DOD/TACOM/NAC/ARC Program R01-1344-18, and by FAA/NSSA Program, R01-1344-48/49.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 69–78, 2004.

There have been various research results of object extraction and tracking. One of the simplest methods is to track difference regions within a pair of consecutive frames [1], and its performance can be improved by using adaptive background generation and subtraction. Based on the assumption of stationary background, Wren et al. proposed a real-time blob tracking algorithm, where the blob can be obtained from object's histogram [5,6].

Shape-based tracking obtains a priori shape information of an object-of-interest, and projects a trained shape onto the closest shape in a certain image frame. This type of methods include contour-based method [7,8,9], active shape model (ASM) [10], state-space sampling, and condensation algorithm [9]. Although the existing shape-based algorithms can commonly deal with partial occlusion, they exhibit several serious problems in the practical application, such as (i) a priori training of the shape of a target object and (ii) iterative modelling procedure for convergence. The first problem hinders the original shape-based method from being applied to tracking objects of unpredictable shapes. The second problem becomes a major bottleneck for real-time implementation.

This paper presents a non-prior training active feature model (NPT-AFM) that generates training shapes in real-time without pre-processing. The proposed AFM can track a deformable object by using a greatly reduced number of feature points rather than taking the entire shape. The NPT-AFM algorithm extracts an object using motion segmentation, and determines feature points inside the object. Such feature points tend to approach toward strong edge or boundary of an object. Selected feature points in the next frame are predicted by optical flow. If a feature point is missing or failed in tracking, an additional compensation process restores it.

In summary major contribution of the proposed NPT-AFM algorithm is twofold: (i) real-time implementation framework obtained by removing a prior training process and (ii) AFM-based occlusion handling using a significantly reduced number of feature points.

The remaining part of this paper is organized as follows. In Section 2, an overview of the proposed tracking framework is given. In Section 3, optical flow-based tracking of feature points is presented. In Section 4, the NPT-AFM-based tracking algorithm for occlusion handling is proposed. Experimental results are provided in Section 5, and Section 6 concludes the paper.

2 Overview of the Feature-Based Tracking Framework

The proposed feature-based tracking algorithm is shown as a form of flowchart in Fig. 1. The dotted box represents the real-time feature tracking, prediction, and correction processes from the tth frame to the $t + 1$st frame. In the object segmentation step we extract an objet based on motion direction by using motion-based segmentation and labeling.

We classify object's movement into four directions, extract suitable feature points for tracking, and predict the corresponding feature points in the next frame. A missing feature point during the tracking process is checked and re-

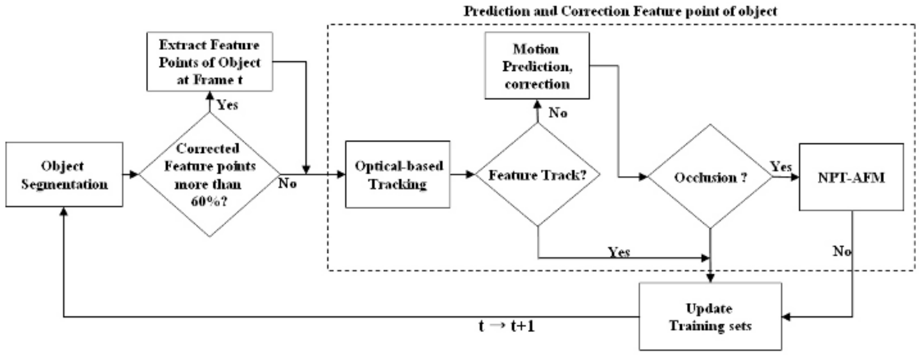

Prediction and Correction Feature point of object
Extract Feature Points of Object at Frame t
Yes
Object Segmentation
Corrected Feature points more than 60%?
No
Optical-based Tracking
Motion Prediction, correction
No
Feature Track?
Occlusion ?
Yes
NPT-AFM
Yes
No
Update Training sets
t → t+1

Fig. 1. The proposed optical flow-based tracking algorithm

stored. If over 60% of feature points are restored, we decided the set of feature points are not proper for tracking and redefine new set of points. We detect occlusion by using labeling information and motion direction, and the NPT-AFM process, which updates training sets at each frame up to 70, restores the entire shape from the occluded input.

The advantages of the proposed tracking algorithm can be summarized as: (i) It can track both rigid and deformable objects without a priori training process and update of the training set at each frame enables real-time, robust tracking. (ii) It is robust against object's sudden motion because both motion direction and feature points are tracked at the same time. (iii) Its tracking performance is not degraded even with complicated background because feature points are assigned inside the object near boundary. (iv) It contains NPT-AFM procedure that can handle partial occlusion in real-time.

3 Optical Flow-Based Tracking of Feature Points

The proposed algorithm tracks feature points based on optical flow. A missing feature point during the tracking is restored by using both temporal and spatial information inside the predicted region.

3.1 Feature Point Initialization and Extraction

We extract motion from a video sequence and segment regions based on the direction of motion using Lucas-Kanade's optical flow method [11]. Due to the nature of optical flow, an extracted region has noise and holes, which are removed by morphological operations.

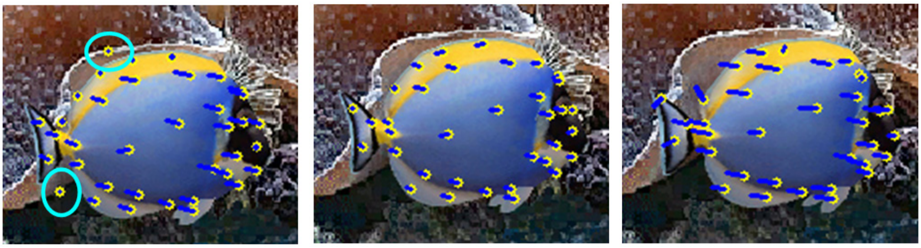
After segmentation of an object from background, we extract a set of feature points inside the object by using Shi-Tomasi's feature tracking algorithm [13]. The corresponding location of each feature point in the following frame is predicted by using optical flow. These procedures are summarized in the following algorithm.

1. Preprocess the region-of-interest using a Gaussian lowpass filter.
2. Compute the deformation matrix using directional derivatives at each pixel in the region as
$$D = \begin{bmatrix} d_{xx} & d_{xy} \\ d_{yx} & d_{yy} \end{bmatrix} \tag{1}$$
where, for example, d_{xx} represents the 2nd order derivative in the x direction.
3. Compute eigenvalues of the deformation matrix, and perform non-maxima suppression. In this work we assume that local minima exist in the 5×5 neighborhood.
4. Discard eigenvalues that is smaller than a pre-specified threshold, and discard the predicted feature points that does not satisfy the threshold distance.
5. Predict the corresponding set of feature points in the next frame using optical flow.

Due to the nature of motion estimation, motion-based segmentation usally becomes a little bit larger than the real object, which results in false extraction of feature points outside the object. These outside feature points are removed by considering the distance between predicted feature points given in

$$d = \sum_{t=1}^{N} \sum_{i=1}^{M} \sqrt{({x_i}^{t+1} - {x_i}^{t})^2 - ({y_i}^{t+1} - {y_i}^{t})^2} < t_n, \tag{2}$$

where t represents the number of frames, and i the number of feature points. The results of outside point removal are shown in Fig. 2.



(a) The 2nd frame (b) The 4th frame (c) The 7th frame

Fig. 2. Results of outside feature point removal (Two outside feature points highlighted by circles in (a) are removed in (b) and (c).)

3.2 Feature Point Prediction and Correction

In many real-time, continuous video tracking applications, the feature-based tracking algorithm fails due to the following reasons: (i) self or partial occlusions of the object and (ii) feature points on or outside the boundary of the object, which are affected by changing background.

In order to deal with the tracking failure, we should correct the erroneously predicted feature points by using the location of the previous feature points and inter-pixel relationship between the predicted points. Here we summarize the prediction algorithm proposed in [12].

1. *Temporal Prediction*: Let the location of a feature block at frame t, which was not tracked to frame $t+1$, be $\underline{v}_i{}^t, i \in \{i, ..., M_t\}$. Its location is predicted using the average of its motion vectors in the previous K frames as
$$\hat{\underline{v}}_i^{t+1} = \hat{\underline{v}}_i^t + \frac{1}{K}\sum_{k=0}^{K-1} \underline{m}_i{}^{t-k}, \tag{3}$$
where $\underline{m}_i{}^t = \underline{v}_i{}^t - \underline{v}_i{}^{t-1}$ denotes the motion vector of feature block i at frame t, and K represents the number of frames for motion averaging . The parameter K may be adjusted depending on the activity present in the scene.
2. *Spatial Prediction*: We can correct the erroneous prediction by replacing with the average motion vector of successfully predicted feature points.
3. *Re-Investigation of The Predicted Feature Point*: Assign a region including the predicted-corrected feature point. If a feature point is extracted in the next frame, it is updated as a new feature point. If more than 60% feature points are predicted, feature extraction is repeated.

The temporal prediction is suitable for deformable objects while the spatial prediction is good for non-deformable objects. Both temporal and spatial prediction results can also be combined with proper weights. In this work, we used $K = 7$ for temporal prediction.

4 NPT-AFM for Robust Tracking

The most popular approach to tracking 2D deformable objects is to use the object's boundary. ASM-based tracking falls into this category. ASM can analyze and synthesize a priori trained shape of an object even if the input is noisy or occluded [14]. On the other hand, a priori generation of training sets and iterative convergence prevent the ASM from being used for real-time, robust tracking. We propose a real-time updating method of the training set instead of off-line preprocessing, and also modify the ASM by using only a few feature points instead of the entire landmark points. NPT-AFM refers to the proposed real-time efficient modeling method.

4.1 Landmark Point Assignment Using Feature Points and AFM

The existing ASM algorithm manually assigns landmark point on the object's boundary to make a training set [14]. A good landmark point has balanced distance between adjacent landmark points and resides on either high-curvature or 'T' junction position. A good feature point, however, has a different requirement from that of a good landmark point. In other words, a feature point is recommended to locate inside the object because a feature point on the boundary of the object easily fails in optical flow or block matching-based tracking [15] due to the effect of changing, complicated background.

Consider n feature points from an element shape in the training set. We update this training set at each frame of input video, and at the same time align the shape onto the image coordinate using Procrustes analysis [16]. In this work the training set has 70 element shapes. Given a set of feature points the input feature can be modeled by using principal component analysis (PCA).

In order to track the target object we have to find the best feature-based landmark points which match the object and the model. In each iteration the feature-based landmark points selected by PCA algorithm are relocated to new position by local feature fitting. The local feature fitting algorithm uses a block-based correlation between the object and the model. The best parameters that represent the optimal location of feature points of the object, can be obtained by matching the feature points in the training set to those of the real image. Here the existing block matching algorithms can be used for the block-based correlation. Figure 3 shows the result of optical flow-based model fitting with 51 training sets.

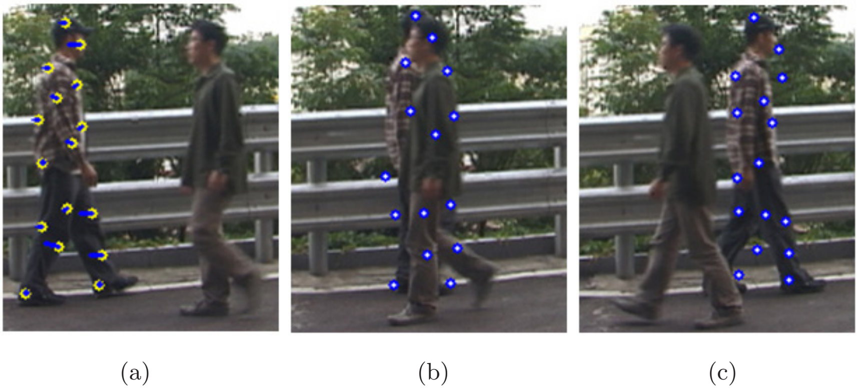


Fig. 3. Model fitting procedure of NPT-AFM: (a) optical flow-based feature tracking at the 40th frame, (b) model fitting at the 74th frame, and (c) model fitting at the 92nd frame.

4.2 Reconstruction of Feature Model and Occlusion Handling

In spite of theoretical completeness of the AFM algorithm,a feature model obtained from the local feature fitting step does not always match the real object because it has been constructed using a training set of features in the previous frame. A few mismatches between the feature model and the real object can be found in Fig. 3.

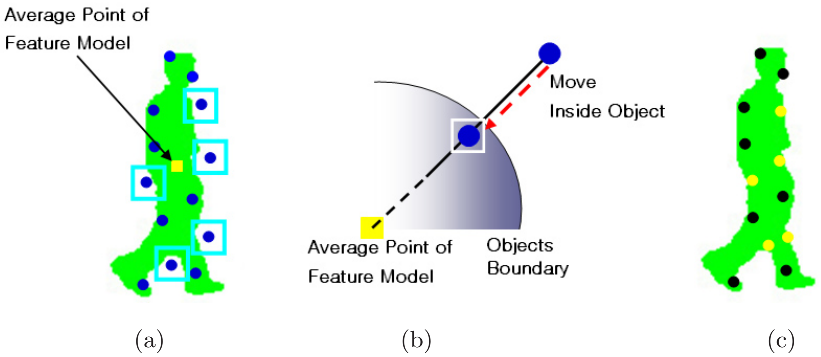


Fig. 4. Reconstruction of the feature model: (a) feature model fitting result, (b) relocation of an outside feature point for feature reconstruction, and (c) result of feature reconstruction.

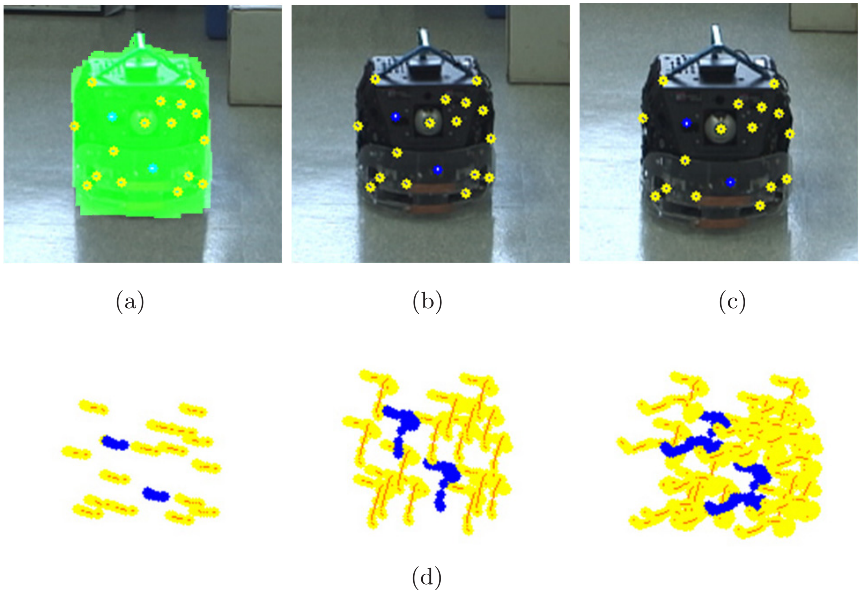
The proposed feature reconstruction algorithm move an outside feature point toward the average position of all feasible feature points, which means feature points inside the object. While moving the outside feature point, we search the best path among three directions toward the average position. If the number of outside feature points is more than 60% of the total feature points, the feature extraction process is repeated. The feature reconstruction process is depicted in Fig. 4

In addition to reconstructing feature model, occlusion handling is another important function in a realistic tracking algorithm. The proposed NPT-AFM based occlusion handling algorithm first detects occlusion if the labeling region is 1.6 times lager than the original labeling region. The decision is made with additional information such as motion direction and size in the correspondingly labeled object region. If an occlusion is detected, we preserve the previous labeling information to keep multiple object's feature models separately. After handling the occlusion, the feature model should be reconstructed every time. This reconstruction process is performed the size of labeled region is between $0.8L$ and $1.2L$, where L represents the original size of the labeled region.

5 Experimental Results

We used 320 by 240, indoor and outdoor video sequences to test tracking both rigid and deformable objects. In most experimental images bright (yellow) circles represent successfully tracked feature points while dark (blue) does ones represent corrected feature points.

For rigid object tracking, we captured an indoor robot video sequence using a Pelco Spectra pan-tilt-zoom (PTZ) camera. We applied the proposed tracking algorithm while the PTZ camera does not move. Once the camera view has changed, we received the relative coordinate information from the camera, and restarted tracking in the compensated image coordinate. The result of tracking is shown in Fig. 5.


(a) (b) (c)

(d)

Fig. 5. Feature tracking of a rigid object and the resulting trajectory: (a) motion-based segmentation result of the 3rd frame, (b) the 10th frame, (c) the 34th frame, and (d) the corresponding trajectory of each frame.

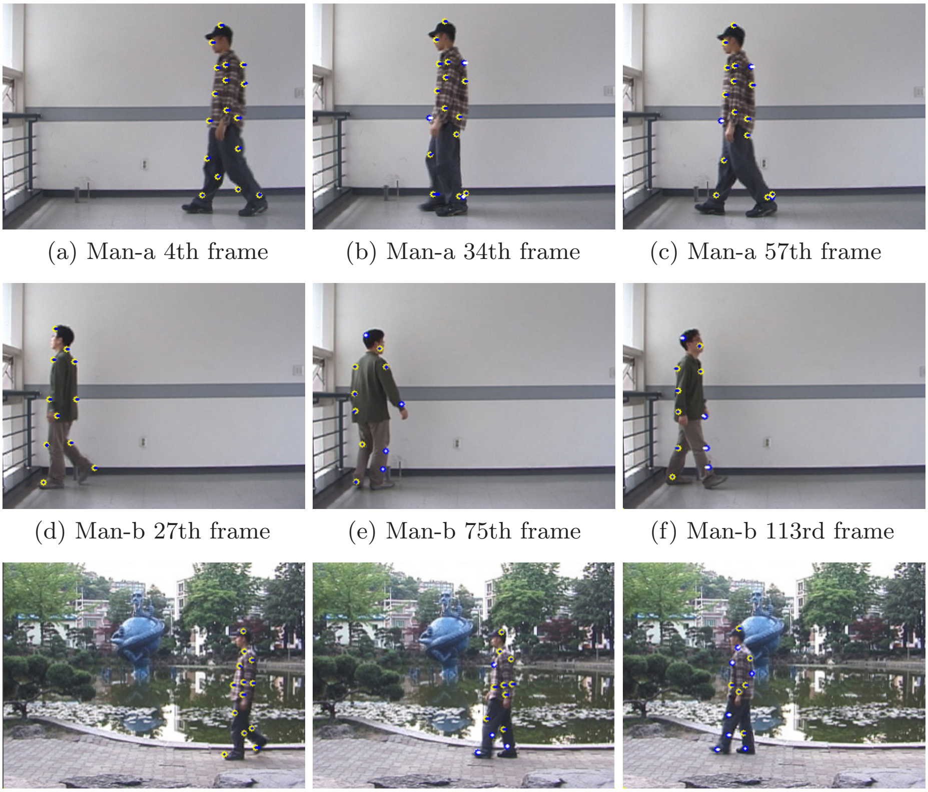
For deformable object tracking, we captured indoor and outdoor human sequences using SONY 3CCD DC-393 color video camera with auto-iris function. Tracking results using the proposed algorithm are shown in Fig .6. Predicted feature points are classified into two classes: successful and reconstructed, which are separately displayed in Fig. 6.

In order to track a deformable object under occlusion, we applied the proposed NPT-AFM-based tracking algorithm. Results of occlusion handling by the proposed NPT-AFM are shown in Fig. 7. By using the NPT-AFM, the proposed tracking algorithm could successfully track an object with occlusion up to 85%.

6 Conclusions

We presented a novel method for tracking both rigid and deformable objects in video sequences. The proposed tracking algorithm segments object's region based on motion, extracts feature points, predicts the corresponding feature points in the next frame using optical flow, corrects and reconstructs incorrectly predicted feature points, and finally applies NPT-AFM to handle occlusion problems.

NPT-AFM, which is the major contribution of this paper, removes the off-line, preprocessing step for generating a priori training set. The training set used for model fitting can be updated at each frame to make more robust object's shape under occlude situation. The on-line updating of the training set can realize a real-time, robust tracking system. Experimental results prove that the

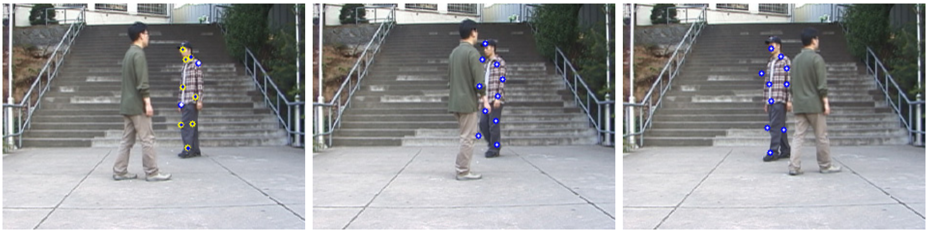


(a) Man-a 4th frame (b) Man-a 34th frame (c) Man-a 57th frame

(d) Man-b 27th frame (e) Man-b 75th frame (f) Man-b 113rd frame

(g) Man-c 5th frame (h) Man-c 31st frame (i) Man-c 43rd frame

Fig. 6. Feature tracking of deformable object in both indoor and outdoor sequences: Bright (yellow) circles represent successfully predicted feature points while dark (blue) circles represent corrected, reconstructed points.



(a) 106th frame (b) 125th frame (c) 165nd frame

Fig. 7. Occlusion handling results using the proposed NPT-AFM algorithm

proposed algorithm can track both rigid and deformable objects under various conditions, and it can also track the object-of-interest with partial occlusion and complicated background.

References

1. Haritaoglu, I., Harwood, D., Davis, L.: W-4: Real-Time Surveillance of People and Their Activities. IEEE Trans. On Pattern Analysis and Machine Intelligence (2000) 809–830
2. McKenna, S., Raja, Y., Gong, S.: Tracking Contour Objects Using Adaptive Mixture Models. Image and Vision Computing (1999) 225–231
3. Plankers, R., Fua, P.: Tracking and Modeling People in Video Sequences. Computer Vision and Image Understanding (2001) 285–302
4. Comaniciu, D., Ramesh, V., Meer, P.: Kernal-Based Object Tracking. IEEE Trans. On Pattern Analysis and Machine Intelligence (2003) 564–577
5. Wren, C., Azerbeyejani, A., Darrel, T., Pentland, A.: Pfinder: Real-Time Tracking of The Human Body. IEEE Trans. Pattern Analysis and Machine Intelligence (1997) 780–785
6. Comaniciu, D., Ramesh, V., Meer, P.: Real-Time tracking of non-rigid objects using mean shift. Proc. IEEE Int. Conf. Computer Vision, Pattern Recognition (2000) 142–149
7. Baumberg, A.: Learning Deformable Models for Tracking Human Motion. Ph.D. Dissertation, School of Comput. Studies (1995)
8. Kass, M., Witkin, A., Terzopoulos, D.: Snake, Active Contour Models. International Journal of Computer Vision (1988) 321–331
9. Blake, A., Isard, M.: Active Contours. Springer, London, England (1998)
10. Cootes, T., Cooper, D., Taylor, C., Graham, J.: Active Shape Models - Their Training and Application. Comput. Image and Vision Understanding **61** (1995) 38–59
11. Bruce, D., Lucas, D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In Proc. DARPA Image Understanding Workshop (1981) 121–300
12. Erdem, C. E., Tekalp, A. M., Sankur, B.: Non-Rigid Object Tracking Using Performance Evaluation Measures as Feedback. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (2001) 323–330
13. Shi, J., Tomasi, C.: Good Features to Track. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (1994) 593–600
14. Koschan, A., Kang, S., Paik, J., Abidi, B., Abidi, M.: Color Active Shape Models for Tracking Non-rigid Objects. Pattern Recognition Letters (2003) 1751–1765
15. Gharavi, H., Mills, M.: Block-Matching Motion Estimation Algorithms: New Results. IEEE Trans. Circ. and Syst (1990) 649–651
16. Goodall, C.: Procrustes Method in The Statistical Analysis of Shape. Journal of The Royal Statistical Society B (1991) 285–339

An Immunological Approach to Raising Alarms in Video Surveillance

Lukman Sasmita, Wanquan Liu, and Svetha Venkatesh

Curtin University of Technology
GPO Box U1987, Perth 6001
Western Australia
{sasmital,wanquan,svetha}@cs.curtin.edu.au

Abstract. Inspired by the human immune system, and in particular the negative selection algorithm, we propose a learning mechanism that enables the detection of abnormal activities. Three types of detectors for detecting abnormal activity are developed using negative selection. Tracks gathered by people's movements in a room are used for experimentation and results have shown that the classifier is able to discriminate abnormal from normal activities in terms of both trajectory and time spent at a location.

1 Introduction

Most current systems [1,2,3] that detect abnormal behaviour work by building models of normal behaviour or activities and detecting deviation from these models as a signature of abnormality. In this paper, we explore an alternative method for abnormal activity detection based on biological immune systems. Instead of building models for *normal* behaviours, the immune system develops a set of *abnormal* detectors, by sampling the entire space and choosing detectors that do not conform to normal. Thus, the abnormal detectors are modeled explicitly and may include cases which are rare or unobserved. To explore this concept, we formulate abnormal detectors to find abnormal behaviours for tracking people in spaces. Such abnormalities could include people walking in areas not normally traversed or spending too much time in a given space. We propose three types of detectors generated by negative selection principle and demonstrate their utilisation in abnormal track detection.

The novelty of this paper lies in an alternate model for abnormal activity detection based on the immune system. Unlike current activity classification systems, this system explicitly models the abnormalities instead of the normal aspects of the activities to be recognised.

The rest of the paper is organised as follows: First, a short review on the human immune system is presented in Section 2. An overview of the architecture and the design of the system is explained in Section 3. Section 4 analyses the results of the experiments. Section 5 concludes the discussion.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 79–86, 2004.

2 Preliminary Background

The immune system (IS) is an adaptive, robust and distributed system that continuously maintains the functions of our body (homeostasis) [4]. It is capable of identifying and eliminating our own cancerous cells (*infectious self*) as well as external microorganisms harmful to the body (*infectious non-self*). The IS monitors the body for an almost unlimited variety of infectious cells, known as *non-self* elements, distinguishing them from native cells of the host (*self* elements). These non-self elements include a plethora of viruses, bacteria and other foreign objects which are collectively known as *pathogens*. In addition, the IS is also capable of memorising past infections to mount a more efficient response to further encounters. Immune cells involved in detection and response are collectively known as *lymphocytes*. These lymphocytes become activated in the presence of external entities such as viruses or bacteria. Specifically, the interacting entities are termed *antigens*.

Lymphocyte cells are designed to match external entities not belonging to our bodies. During maturation, immature lymphocyte cells are exposed to self-antigens. If lymphocytes bind to self-antigens, they undergo programmed self-destruction [5]. Cells that survive the maturation period, therefore must *not* bind to self-antigens and consequently bind to only nonself-antigens. This selective process is called negative selection [6].

The collection of lymphocytes as a whole represents the complementary set of our self cells. Trillions of lymphocyte cells exist roaming around our body detecting pathogens [5]. The large number is needed so that the system can model all non-self elements completely. The size of the detector sets thus provide a measure of completeness, as small sets cannot represent a complete abnormality model but on the other hand, limited resources enforce a limit on how large the set should be. A more detailed explanation of the immune system can be found in [7]. [6], [8] and [9] have investigated the use of immunological concepts in network security, specifically in intrusion detection systems (IDS). IDS is a system put in place over a network to detect *misuse* or *anomalies* [9] in the network. The IDS builds an image of self from the normal activities of the network, and treats anomalies as non-self elements.

3 Proposed Approach

3.1 System Architecture

In the proposed surveillance system, image data was acquired by using a top-view camera overlooking a room 7 metres long by 7 metres wide. Background subtraction [2] is initially performed to extract the object and a Kalman filter [10] is then used to track the object. Examples of isolated objects are shown in Figure 2. The center of the bounding rectangle is used as the *observed* position of the object in the real world.

The use of the centre of the bounding box to represent the position of the object results in a noisy observation. Since the object width and height determine

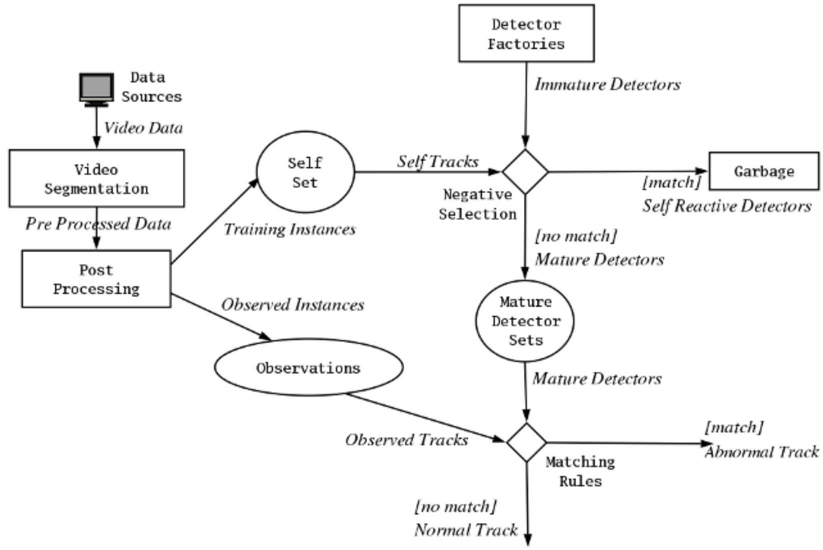


Fig. 1. System Architecture

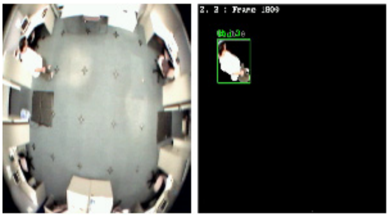


Fig. 2. Raw frame and its isolated foreground

the centre of the box, an object such as a person standing still but moving his/her limbs may change the *observed* positions as the box changes its size. To remedy this problem, an averaging method is used to smooth the position data. The number of observations used to calculate the average is termed the *window size*. This method results in a reduced amount of detail of the observed object but smoothes the observed position data.

The overall architecture of the system is shown in Figure 1. The system is divided into two modules, the *training* module and the *deployment* module. The training module (the top half of Figure 1) involves recording the movements of the objects (or people) in the room over a period of time. These movements form the self set which define the normal behaviours in that room. This self set is then used in the detector generation process.

Detectors are generated randomly by *factories*. These random detectors are then exposed to the self set. If a detector is deemed to match a behaviour in the self set, then it is discarded. Randomly generated detectors that do not match the self set are then admitted into the system as *mature* detectors. This process closely follows the negative selection algorithm in the immune system.

After training, the system is deployed where it will continually survey moving objects in a room. If the behaviour of an object is considered by a detector set to be abnormal, then an alarm is raised.

3.2 Point Detector

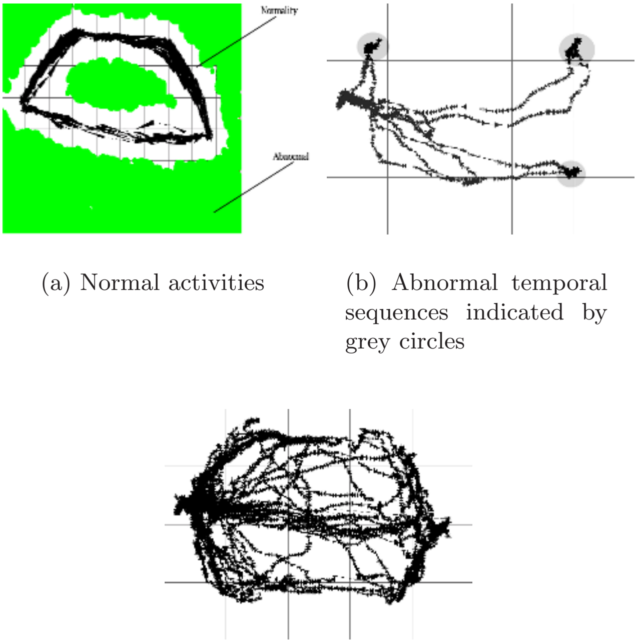
We introduce three types of detectors to detect abnormality. An object that is spatially normal will share the same spatial locality as other normal objects. For example, people traversing through a room will share one or more multiple paths, thus these paths define what is normal in that room. If a path is found to be deviating away from the normal space, then the path is abnormal. A *point detector* is a circle on the observed space. The detector triggers if the object's observed position falls inside the detector's circle. Since a point detector is generated by negative selection, the area bounded by the detector models abnormal space. Therefore, an object that has entered abnormal space, will trigger multiple point detectors and raise an alarm for being spatially abnormal.

A visual representation of the naive point detectors after training can be seen in Figure 3(a). The black arrows represent the tracks of normal people as they move throughout the room. In this case, abnormal space are the areas that are not traversed by the people and are already covered by mature detectors. There is an envelope surrounding each normal track to account for variability in the behaviours.

3.3 Time Detector

Aside from spatial abnormality, a temporal abnormality occurs when a person is stationary in one position for a long time. Point detectors cannot raise an alarm from such behaviours because they do not encode a notion of time. The second type of detector is able to do so and we termed it a *time detector*. The conception and structure of a time detector is similar to a point detector. The main difference is in the matching rule. A time detector has an additional variable *time limit* which maintains a bound as to how long a person can stay inside the circle bounded by that time detector. If the length of time in which a person is stationary exceeds the threshold, then the person is considered to have a temporal abnormality. The length of time that a person has been stationary is given by how many observed positions share the same spatial locality.

Figure 3(b) shows track where people spend a long period of time being stationary in the room. These tracks are normal spatially because they traverse the normal areas shown in Figure 3(a). Although normal in spatial dimension, they are abnormal temporally because of the long duration of stationary period.



(a) Normal activities

(b) Abnormal temporal sequences indicated by grey circles

(c) Abnormal tracks

Fig. 3. Observed tracks in a room

Those periods are shown in grey circles. Figure 3(c) shows a collection of tracks which are abnormal when compared to Figure 3(a).

3.4 Trajectory Detector

The third type of detector attempts to detect an abnormal behaviour that can only be detected by observing the direction of the moving object. We term this detector a *trajectory* detector and its working is similar to the chain coding approach. A vector can be drawn from the object's previous position to the object's current position. The major angle that this vector makes to the *north* vector, is the direction of the object at that time.

The structure of a trajectory detector is a list of angles. The detector is deemed to have found a match if this list of angles matches with a sequence of angles acquired from the object's trajectories.

This detector is affected by the smoothing process described above. Without smoothing, the noisy data causes small but many trajectory changes. By having many trajectory changes, the system becomes sensitive and we expect to have many false alarms. With the addition of smoothing, small changes in trajectories will be removed and therefore we expect to have the rate of false alarms reduced. However, at larger smoothing levels we will expect a drop in the rate of the positive alarms due to the reduced track resolution.

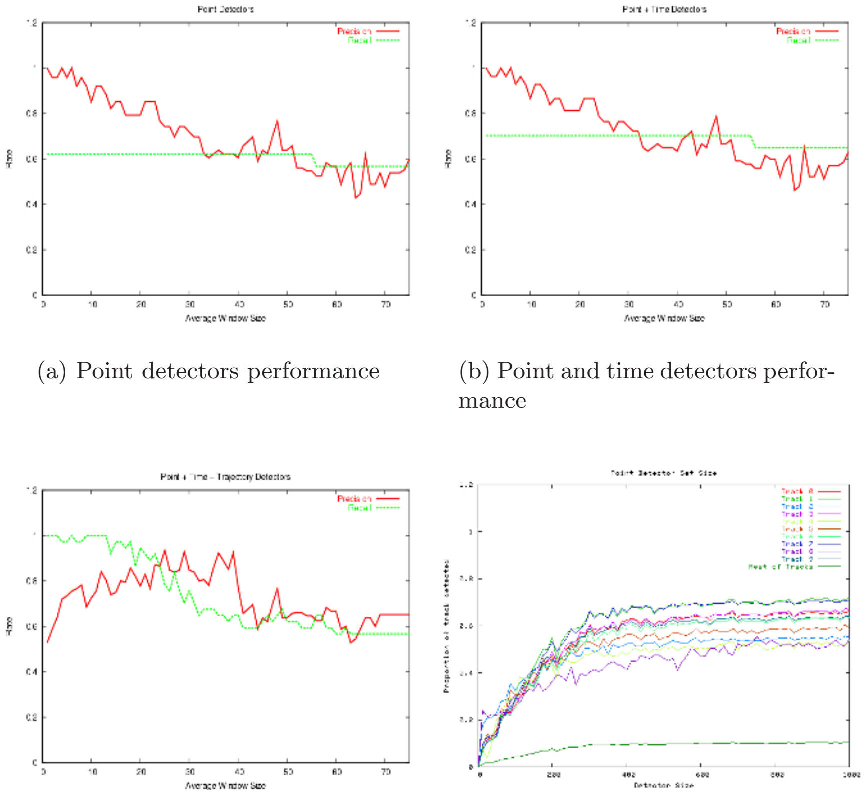
4 Experiments

4.1 Discrimination Capability

A series of recordings were taken which consists of people moving in the room over a certain period of time. The system was first trained with the notion of self, which were the normal activities in the room. Naive detector sets for the three types of detectors were generated according to Figure 1. Next, the system was deployed and exposed to people in the room. In this exposure, 34 sequences or tracks which are abnormal were presented along with 40 normal tracks for which the system should not raise suspicion. The experimentation was done first with the point detectors, and then with time and trajectory detectors. This was to determine the discriminating performance of the system as more complex detectors are added. The experiments were also done with varying window sizes, to test our hypothesis that a small window size will produce many alarms due to trajectory detectors being too sensitive. Figures 4(a), 4(b), 4(c) shows the precision and recall values plotted against the window size. The horizontal axis of the graph indicates increased levels of smoothing. The recall of all detectors declined when the window size is increased due to the decreased detail available.

The graphs in Figure 4(a), 4(b) and 4(c) show that the recall of the system (the thicker line) increases as more complex detectors are added. In Figure 4(b), the recall increased with the added discrimination power of the time detectors. In Figure 4(c), without any smoothing of data the three detectors successfully raised an alarm for every abnormal behaviour (100% recall rate). However, there is a decrease in the recall rates of the three experiments as the level of smoothing is increased. In the worst case, a large smoothing will cause all tracks to become straight. Therefore, what is previously an abnormal track is now normal to the system.

The precision trend among the three graphs also decrease as the smoothing level increases. The system becomes less precise in raising alarms because of the reduction in detail of the track. This reduction may add an element of abnormality to a normal track and *vice versa*. In Figure 4(c), the precision first increases and then decreases. The initial low precision rate in Figure 4(c) is due to trajectory detectors raising false alarms bacause of the noisy data. As smoothing reduces the noise, the rate of false alarms decreases, increasing the precision rate. At some point, the precision rate starts to decrease due to the lost detail as the averaged values no longer reflect the original values. This is inline with our hypothesis that the trajectory detectors are heavily affected by the smoothing process. However, an optimal smoothing window size can be determined from Figure 4(c) and is in the range of 20 to 25. The introduction of trajectory detectors comes with the expense of a smoothing process being used.



(a) Point detectors performance

(b) Point and time detectors performance

(c) Point, time and trajectory detectors performance

(d) Size of point detector set

Fig. 4. Experimental Results

However, the tradeoff of losing the resolution of the observation by smoothing is justified by the increase in the system's discriminating power.

4.2 Size of Detector Set

An experiment was performed to investigate what is the minimum point detector set size that can be used without reducing the system performance. In this experiment, the total area of the space was 250000 $units^2$ and the average area of a point detector was 706 $units^2$ (detector radius of 5 to 25 units), which was only 0.20% of the whole space. In each experiment set, the detector size was increased by 5 units. Next, the system was exposed to 10 abnormal tracks and 40 other normal tracks to test its capability to raise an alarm. Figure 4(d) shows the experimental results. The horizontal axis shows an increasing number of detectors in the set, while the vertical axis represents the proportion of each track that is detected by the point detector set, which is the proportion of the track considered by the set to be abnormal. The system detects a large proportion of track 0 to 9 to be abnormal. The proportion of abnormality detected increases as the detector size becomes larger. However, after a size of 600 the proportion detected stays constant. Therefore, we can safely set the minimum point detector size to be 600, since larger sets do not add to the discrimination power in this data set. This is a very small size considering that a single detector only covers 0.20% of the total space. This stems from the fact that cross reactivity and the multi reactivity of the detectors can provide sufficient detection [7].

5 Conclusions

The paper has shown that applying immune system concepts such as negative selection provides a useful solution for abnormal activity detection. Although only three types of abnormal detectors were examined, more complex detectors which take into consideration the context of the room deserve further investigation and may reduce the rate of false alarms.

References

1. Makris, D., Ellis, T.: Path detection in video surveillance. Image and Vision Computing **20** (2002) 895–903
2. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 747–757
3. Johnsom, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. Image and Vision Computing **14** (1996) 609–615
4. Janeway, A, C., Travers, P.: Immunobiology: The Immune system in Health and Disease. 2nd edn. Garland Publishing, Inc., New York (1996)
5. Hofmeyr, S.A.: An interpretative introduction to the immune system. In Cohen, I., Segel, L., eds.: Design Principles for the Immune System and other Distributed Autonomous Systems. Oxford University Press (2000)
6. Forrest, S., Perelson, A., Allen, L., Cherukuri, R.: A change-detection algorithm inspired by the immune system. [Online] http://citeseer.nj.nec.com/51699.html (1995)
7. de Castro, L.N., Von Zuben, F.J.: Artificial immune systems: Part I: Basic theory and applications. Technical Report 01/99, DCA (1999)
8. Hofmeyr, S.A., Forrest, S.: Architecture for an artificial immune system. Evolutionary Computation **7** (2000) 45–68
9. Puglisi, S.: An immunological approach to network intrusion detection (2002) Honours Thesis, Department of Computer Science, Curtin University of Technology.
10. Kalman, R.: A new approach to linear filtering and prediction problems. In: Transaction of the ASME – Journal of Basic Engineering. (1960) 33–45

Sat-Cam: Personal Satellite Virtual Camera

Hansung Kim[1,2], Itaru Kitahara[1], Kiyoshi Kogure[1], Norihiro Hagita[1], and Kwanghoon Sohn[2]

[1] Intelligent Robotics and Communication Laboratories, ATR, 2-2-2 Hikaridai,
Keihanna Science City, Kyoto, 619-0288 Japan
{kitahara,kogure,hagita}@atr.jp
http://www.irc.atr.jp

[2] Dept. of Electrical and Electronics Eng., Yonsei University 134 Shinchon-dong,
Seodaemun-gu, Seoul, 120-749, Korea
hskim99@diml.yonsei.ac.kr, khsohn@yonsei.ac.kr
http://diml.yonsei.ac.kr

Abstract. We propose and describe a novel video capturing system called Sat-Cam that can observe and record the users' activity from effective viewpoints to negate the effects of unsteady mobile video cameras or switching between a large number of video cameras in a given environment. By using real-time imagery from multiple cameras, our system generates virtual views fixed in relation to the object. The system consists of a capturing and segmentation components, a 3D modeling component, and a rendering component connected over a network. Results indicate that a scene rendered by Sat-Cam provides stable scenes that help viewers understand the activity in real time. By using the 3D modeling technique, an occlusion problem in object tracking is solved, allowing us to generate scenes from any direction and zooming in/out condition.

1 Introduction

With the progress of communication technology, many systems are being developed to share experience or knowledge and interact with other people [1,2,3,4]. As the proverb goes –"a picture is worth a thousand words" – we can understand most of other people's activities by watching their videos. In this paper, we propose a novel video capturing system called Sat-Cam that can observe and record the users' activity from effective viewpoints. Applications of this technology extend to surveillance, remote education or training, telecommunication, and so on.

There are two typical approaches to capturing visual information of working records. As illustrated in Figure 1(left), the first one is to use a head-mounted or mobile camera attached to users [3][4][5][6]. This approach can easily record all activities the user performs with the minimum amount of data (i.e., single video stream). However, when a user, who does not have same context as the captured user, tries to understand his/her experiences by watching the video, the mobile camera may cause the user inconvenience, because the video data is captured from a subjective viewpoint. For example, the sway of the camera

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 87–94, 2004.

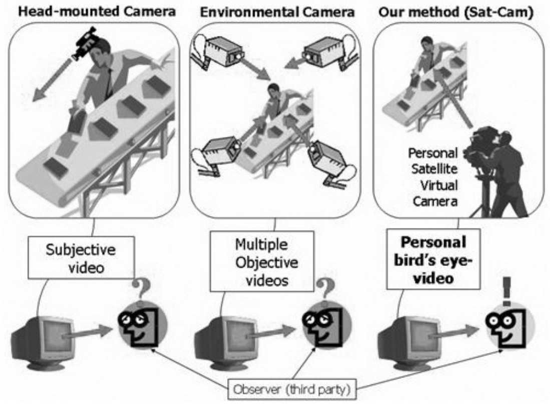


Fig. 1. Approaches to capturing visual information of user's activities

leads to viewers becoming confused; moreover, it is difficult for third parties to understand the scene from the camera since the movement of the user itself cannot be observed.

The second approach, which is illustrated in Figure 1 (center), is to use cameras fixed in an environment. Since the cameras always provide objective and stable visual information, it is easier for third parties to understand them. However, an enormous amount of useless video must be captured to cover the whole area at all times. To determine the best situation for observing the activities, we have to switch between multiple videos. By increasing the number of capturing cameras, this switching-monitoring operation sometimes exceeds a human's processing ability.

To overcome the above problems, we propose the Sat-Cam system, is illustrated in Figure 1 (right), which is a method for capturing a target object from a bird's-eye view. The system generates virtual views fixed relative to the object by reconstructing its 3D model in real time. The scene rendered by Sat-Cam provides stable scenes of minimal area for understanding the user's activity, making it easier for third parties to understand. The proposed system requires higher computational cost than conventional approaches do. Therefore, we applied and proposed efficient segmentation and modeling algorithms for real time processing.

In the next section, we provide an overview of the Sat-Cam system. Section 3 then describes the detailed algorithms used in the system. Section 4 shows the experimental set up and result, and finally, we draw conclusions in Section 5.

2 Sat-Cam System

Figure 2 shows the concept of Sat-Cam. CV (Computer Vision)-based 3D video display systems have become feasible with the recent progress in computer and video technologies, and indeed several systems have been developed[7][8]. As "Sat-Cam" stands for "Satellite Camera," the system aims to capture the visual

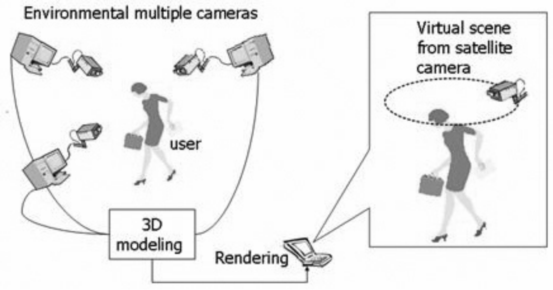

Environmental multiple cameras
user
3D modeling
Rendering
Virtual scene from satellite camera

Fig. 2. Overview of Sat-Cam System

information of working records by a virtual camera that orbits the target user, employing a 3D video processing technique. Since the virtual camera always tags along with the user, it can record all activities the user performs as a single video stream. When Sat-Cam's point of view is set to look down on the target space like a satellite, the captured video can be easily understood by third parties. One of the most important features of this system is that it works in real time. If we generate the Sat-cam video in the post-process, it is necessary to record enormous amounts of environmental video data.

3 Algorithms of the Proposed Method

This section describes in detail the algorithms used by the system. The system comprises three sub-systems: object segmentation in capturing PCs, 3D modeling in a 3D modeling server, and rendering virtual views in a rendering PC.

When target objects are captured by cameras, each capturing PC segments the objects and transmits the segmented masks to a 3D modeling server. The modeling server generates 3D models of the objects from the gathered masks, and tracks each object in a sequence of 3D models scenes. The 3D model, object ID (identification) and 3D position of the objects, are sent to a rendering PC via a network, and finally, the rendering PC generates a video at the designated point of view with the 3D model and texture information from cameras.

3.1 Object Segmentation

Real-time object segmentation is one of the most important components of the proposed system, since the performance of the segmentation decides the quality of the final 3D model.

We realized the object segmentation of color images based on Chien's [9] and Kumars' [10] algorithms using background subtraction and inter-frame differences. At first, the background is modeled with minimum and maximum intensities of the input images which are low-pass filtered to eliminate noise. Then, the frame difference mask is calculated by thresholding the difference between two consecutive frames. In the third step, an initial object mask is constructed from the frame difference and background difference masks by the OR process. Forth,

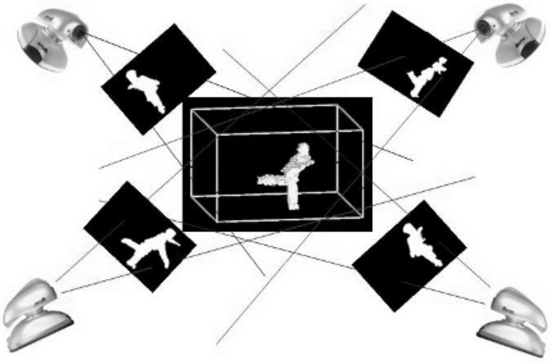


Fig. 3. 3D modeling with the shape-from-silhouette technique

we refine the initial mask by a closing process and eliminate small regions with a region-growing technique. Finally, in order to smoothen the objects' boundaries and to eliminate holes inside the objects, we applied Kumar's profile extraction technique[10] from all quarters.

If the target space has poor or unstable illumination, thermal cameras can be used. In this case, the segmentation process is much simpler than for color images since a human object is brighter than the background. However, using thermal cameras will increase the expenses of the system, so it can be used as an option. We make an initial mask by thresholding with the intra-variance from the mean of the thermal scene.

The final segmented mask is converted into binary code and transmitted to the modeling server via UDP (User Datagram Protocol).

3.2 3D Modeling

The transmitted binary segmented images from the capturing PCs are used to reconstruct a 3D model of the objects. The modeling PC knows projection matrices of all cameras because they were calibrated in advance.

We use the shape-from-silhouette technique to reconstruct a 3D model as shown in Figure 3 [11]. The check points $M(X, Y, X)$ in 3D spaces are projected onto multiple images I_n with the following equation, where P_n is a projection matrix of a camera C_n;

$$(u, v, 1)^T = P_n(X, Y, Z, 1)^T \tag{1}$$

If all projected points of M are included in the foreground region of multiple images, we select the point as inside voxel of an object.

Testing all points in a 3D model is, however, a very time-consuming process and results in heavy data. Therefore, we used an octree data structure for modeling. For each voxel of a level, 27 points (i.e., each corner and the centers of edges, faces and a cube) are tested. If all checking points are either included in or excluded from an object, the voxel is assigned as a full or empty voxel, respectively. Otherwise, the voxel splits into eight sub-voxels and is tested again at the

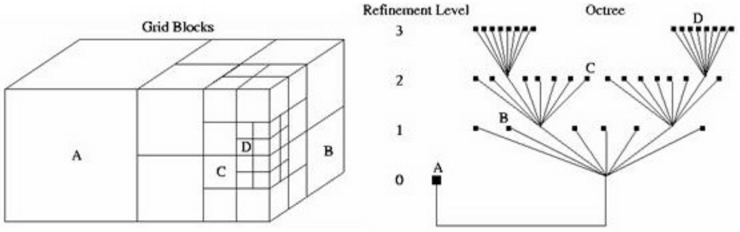


Fig. 4. Octree structure

next refinement level. Figure 4 shows the structure of the octree. Its structure dramatically reduces the modeling speed and the amount of data.

After the modeling process, the server performs object tracking in the model. It is very difficult, however, to track in the 3D model in real time; therefore, we perform the tracking in the 2D plane.

We assume that ordinary objects (human) have a constant height (e.g., 170 cm), and extract a 2D plane model by slicing the 3D model at a lower height (e.g., 120 cm). Then, we grow and label the regions on the plane model. By tracking the center of each labeled region in a series of model frames, we can identify and track each object.

Finally, modeling parameters, 3D positions of objects with ID numbers, and node information of an octree model are transmitted to the rendering part.

3.3 Virtual View Rendering

In the rendering part, the received 3D model is reconstructed and the virtual view of Sat-cam is synthesized. When the octree information is received, it reconstructs the 3D model by decoding the node information and inserts the model at the correct position in 3D space. The transmitted data from the modeling server also includes 3D position information and object IDs of any objects. The rendering PC requests texture information an objects to capturing PCs, and performs texture mapping onto the reconstructed 3D model.

However, the resolution of our 3D model is not sufficient for a simple (1-on-1) texture mapping method because we place real-time processing ahead of reconstructing a fine 3D model so that the octree method describes the 3D model with several levels of resolution. Our system employs the “Projective Texture Mapping Method” to solve this problem [12]. This mapping method projects the texture image onto the 3D objects as if a slide projector. Consequently, the resolution of the texture image is retained during the texture mapping process, regardless of the resolution and shape of the mapped 3D model. Moreover, this method is implemented as OpenGL functional libraries; it is possible to take advantage of a high-speed graphic accelerator. By merging the working space (background), which is modeled in advance, a complete 3D model of the working space and object is reconstructed.

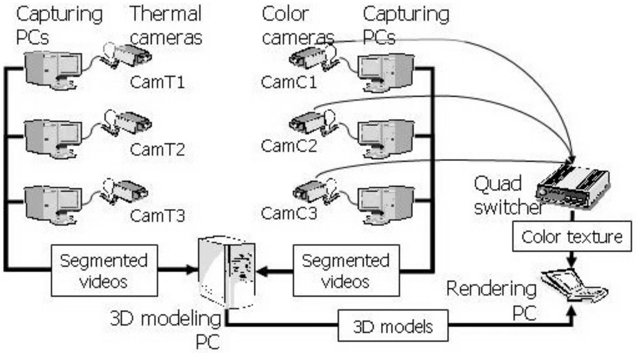


Fig. 5. Configuration of our pilot Sat-Cam System

Finally, the rendering PC generates scenes at the viewpoint requested by a user. This system provides the following two modes to control the viewpoint of a virtual camera.

Tracking mode: In this mode, the virtual camera observes the object from a position above and behind the user. The direction of the object should be known in order to control the pan and the tilt value of the virtual camera. We assumed that the direction toward which an object moves is the same to its front direction. The direction of movement is estimated by tracking a global path of objects from the movements in the previous consecutive frames. While the object is moving, this controlling mode is applied.

Orbiting mode: We can make the virtual camera go around the object like a satellite when the target object stops in one position. Thus, orbiting mode makes it possible to observe the blind (self-occluded) spots.

4 Implementation of Sat-Cam System

As shown in Figure 5, we have implemented a distributed system using eight PCs and six calibrated cameras (three SONY EV-100 color cameras and three AVIO IR-30 thermal cameras). The systems are realized with commercially available hardware. Six portable Celeron 800-MHz PCs are used to capture the video streams and segment objects. The segmented information is sent via UDP over a 100-Mb/s network to the modeling PC. The modeling and rendering PCs have Pentium-4 CPUs, and GeForce-4 FX5200 and Quadro FX1000 graphic accelerators, respectively.

The segmentation information from each camera has a resolution of 180×120 and the 3D space has a resolution of $256 \times 128 \times 256$ on a 2cm voxel grid. It covers an area of about 25m^2 areas and 2.5m height. We set up some parameters in the segmentation process as follows: 5 as a threshold for frame difference, 100 for the smallest region size, 5 for elasticity to make silhouette.

Table 1 shows a run-time analysis with our algorithm. The times listed are average times for a single target to exist in a working space. The bottleneck in the system is the 3D modeling process, since it is performed in 3D space.

Table 1. Run-time analysis (msec)

Segmentation		3D Modeling		Rendering	
Function	Time	Function	Time	Function	Time
Capturing	66.15	Receiving	0.23	Receiving	0.31
Segmentation	2.44	Initialization	32.70	Capturing	
Closing	1.35	3D Modeling	85.29	Texture	67.02
Elimination	5.28	Labeling	1.76	Rendering	85.70
Silhouette	7.48	Tracking	0.17	Flushing	0.04
Transmission	0.95	Transmission	1.84		
Total	83.65ms	Total	121.99ms	Total	103.07ms
Frame/sec	11.95f/s	Frame/sec	8.20f/s	Frame/sec	9.70f/s

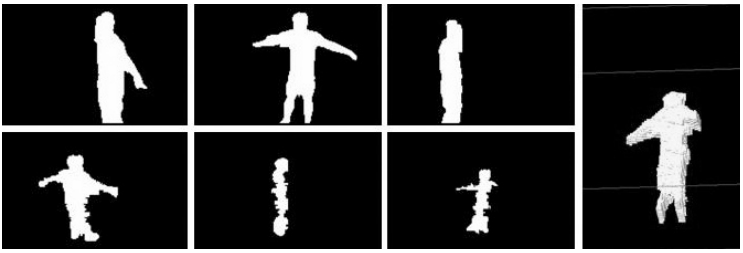


Fig. 6. 3D modeling results: the upper row is the result of segmentation process with using thermal cameras; the lower row is the result with using color cameras. The reconstructed 3D model is shown in the rightmost cell.

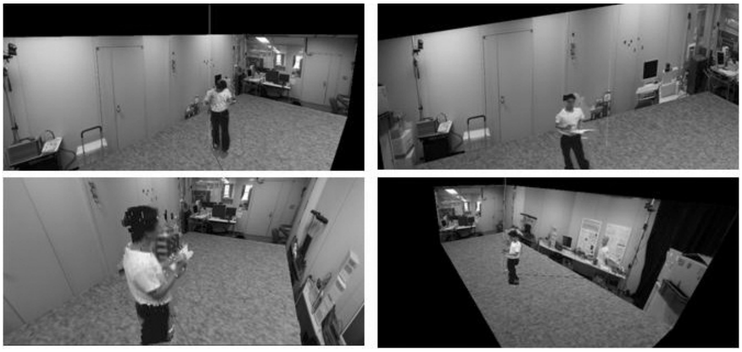


Fig. 7. Rendered scenes by Sat-Cam

However, the frame rate of the whole system shows about 10 frames per second, though it depends on the complexity of the objects.

Figure 6 shows snapshots of segmented images and a constructed 3D model. Generally, thermal cameras provide more reliable segmented information. Therefore, we assigned higher priority to the information from thermal cameras, this priority can be adjusted since their reliability may decrease in the case where people put on warm clothes. The rendered scene from a rendering PC is shown in Figure 7.

5 Conclusions and Future Works

We proposed a novel video capturing system called Sat-Cam that can observe and record the users' activities from effective viewpoints to negate the effects of unsteady mobile video cameras or switching between a large number of videos in a given environment. By using real-time imagery from multiple cameras, the proposed system generates virtual views fixed in relation to the object. The system provides stable scenes that enable viewers to understand the user's activity in real-time. However, we should solve several problems on realizing Sat-Cam, e.g. how to keep the image quality of rendered objects (especially when the Sat-Cam comes in the middle of two real cameras), how to cope with invisible areas when viewers try to see the area, etc. In future works, we will consider evaluating the Sat-Cam's usefulness for sharing life-log information between users.

Acknowledgements. This research was supported in part by the National Institute of Information and Communications Technology.

References

1. Xu, L.Q., Lei, B.J., Hendriks, E.: Computer Vision for a 3-D Visualisation and Telepresence Collaborative Working Environment. BT Technology Journal, Vol. 20, No. 1 (2002) 64–74
2. Kuwahara, N., Noma, H., Kogure, K., Hagita, N., Tetutani, N., Iseki, H.: Wearable Auto-Event-Recording System for Medical Nursing. Proc. INTERACT'03 (2003) 805–808
3. Kawashima, T., Nagasaki, T., Toda, M., Morita, S.: Information Summary Mechanism for Episode Recording to Support Human Memory. Proc. PRU, (2002) 49–56
4. Nakamura, Y., Ohde, J. Ohta, Y.: Structuring Personal Activity Records based on Attention - Analyzing Videos from Head-mounted Camera. Proc. 15th ICPR, (2000) 220–223
5. Yamazoe, H., Utsumi, A., Tetsutani, N., Yachida. M.: Vision-based Human Motion Tracking using Head-mounted Cameras and Fixed Cameras for Interaction Analysis. Proc. ACCV, Vol.2 (2004) 682–687
6. Ohta, Y., Sugaya, Y., Igarashi, H., Ohtsuki, T., Taguchi, K.: Share-Z: Client/Server Depth Sensing for See-Through Head-Mounted Displays, PRESENCE, Vol. 11, No. 2 (2002) 176–188
7. Kanade,T., Rander, P.W., Narayanan, P.J.: Virtualized Reality: Constructing Virtual Worlds from Real Scenes. IEEE Multimedia, Vol.4, No.1, (1997) 34–47
8. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-Based Visual Hulls, ACM SIGGRAPH 2000, (2000) 369–374
9. Chien, S.I., Ma, S.Y, Chen, L.G.: Efficient Moving Object Segmentation Algorithm using Background Registration Technique. IEEE Trans. on CSVT, Vol. 12, No. 7 (2002) 577–586
10. Kumar, P., Sengupta, K., Ranganath, S.: Real Time Detection and Recognition of Human Profiles using Inexpensive Desktop Cameras. Proc. 15th ICPR, Vol. 1 (2000) 1096–1099
11. Kitahara, I., Ohta, Y.: Scalable 3D Representation for 3D Video Display in a Large-scale Space. Proc. IEEE Virtual Reality (2003) 45–52
12. Everitt. C.: Projective Texture Mapping. NVIDIA SDK White Paper

A Linear Approximation Based Method for Noise-Robust and Illumination-Invariant Image Change Detection

Bin Gao[2]*, Tie-Yan Liu[1], Qian-Sheng Cheng[2], and Wei-Ying Ma[1]

[1] Microsoft Research Asia, No.49 Zhichun Road, Haidian District,
Beijing 100080, P. R. China, +86-10-62617711
{t-tyliu,wyma}@microsoft.com

[2] LMAM, Department of Information Science, School of Mathematical Sciences,
Peking University, Beijing 100871, P.R. China, +86-10-51637832
gaobin@math.pku.edu.cn, qcheng@pku.edu.cn

Abstract. Image change detection plays a very important role in real-time video surveillance systems. To deal with the illumination, a category of linear algebra based algorithms were designed in the literature. They have been proved to be effective for surveillance environment with lighting and shadowing. In practice, other than illumination, the detecting process is also influenced by the noises of cameras and reflections. In this paper, analysis is made systemically on the existing linear algebra detectors, showing their intrinsic weakness in case of noises. In order to get less sensitive to noises, a novel method is proposed based on the technique of linear approximation. Theoretical and experimental analysis both show its robustness and high performance for noisy image change detection.

Keywords: Video surveillance, change detection, linear algebra.

1 Introduction

It has been a long-time studied topic in computer vision to detect changes in images taken at the same scene but at different times by a static camera. It serves as the basis of a large number of applications including video surveillance, medical diagnosis, civil infrastructure and so forth. However, varying illumination conditions, shadows, reflections and the noises of cameras make the veracious and robust detection a hard work. If the algorithm is not well designed, these influences will lead to false alarms, even when there are no changes at all.

In the past twenty years, many approaches have been proposed for image change detection. In [1], the calculation of changes between two images was performed on predefined sliding windows over each pixel. A statistical description of the ensemble of pixels was given and the decision about the change was made by statistical hypothesis test. Similar methods based on likelihood ratio tests

* This work was performed at Microsoft Research Asia.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 95–102, 2004.

were developed in [2] and [3]. Hsu et al. [4] fit the intensity values in each sliding window to a polynomial function of the pixel coordinates, and compared different likelihood tests using constant, linear and quadratic models. The polynomial model used in [4] was further extended to be illumination-invariant by introducing the partial derivatives on the quadratic model [5]. Besides, many other methods were also proposed in the literature. Among these methods, a series of linear algebra based approaches [6,7] attracted great attention because they are illumination-invariant. This is an important improvement as compared to the previous solutions. However, as analyzed in the following sections of this paper, the performance of these algorithms drops significantly in cases of noises. As we know, in practical environments, there always exist additive noises in the images. So it is a must for a practical change detection algorithm to be robust to noises. To tackle this problem, we proposed a novel method based on linear approximation. Theoretical analysis and experiments both show that this method outperforms the previous linear algebra based methods greatly in case of noises.

The rest of this paper is organized as follows. In Section 2 the general theories of linear algebra based image change detectors are described and the noise analysis is made. In Section 3 the new method is presented. Experimental results are discussed in Section 4. Then the conclusion remark is drawn in the last section.

2 Noise Analysis on Existing Linear Algebra Methods

As mentioned above, there have been several change detection techniques employing linear algebra in the literature. Almost all of them are closely related to the shading model [5] and working with the concept of linear dependence. The basic idea is that if there is no change, the pixel intensities in the current and the reference images should be linear dependent in spite of the illumination. In this sense, when there is no noise, different algorithms are equivalent. However, the situation may change when noises exist. To make it clear, in this section, we will focus on the theoretical analysis of such algorithms' performance by taking noise into consideration.

2.1 Shading Model

In the shading model, the intensity $F(x, y)$ of a pixel(x, y) can be modelled as the product of the illumination $I(x, y)$ from the light sources and the reflectance coefficient $R(x, y)$ of the object surface [8]:

$$F(x, y) = I(x, y)R(x, y). \tag{1}$$

Such a model covers most of the influences mentioned in the introduction: illumination, shadowing and reflection. By applying it to both the current and the reference images, we have,

$$F_r(x, y) = I_r(x, y)R_r(x, y), \quad F_c(x, y) = I_c(x, y)R_c(x, y), \tag{2}$$

where r and c represent the reference and current images respectively. Since the reflectance coefficient depends only on the physical structure of the object surface, $R_c(x,y)$ and $R_r(x,y)$ should be equal for the same pixel. Then we have:

$$F_c(x,y)/F_r(x,y) = I_c(x,y)/I_r(x,y) = k(x,y). \tag{3}$$

When working on a small area, it is reasonable to approximate that the illuminations $I_c(x,y)$ and $I_r(x,y)$ are independent of the pixel positions (x,y). Accordingly $k(x,y)$ will become a constant. In other words, $F_c(x,y)$ and $F_r(x,y)$ are linear dependent. That is just the key point of the shading model.

2.2 Linear Dependence Models

In the linear dependence detector (LDD) proposed by Durucan and Ebrahimi [6], a vector model as illustrated in Fig. 1 was used: a center pixel and its neighbors form a sliding window, and then the center pixel is represented by a vector made up of all pixels in the window. The windows usually take a size of 3 × 3, 5 × 5, 7 × 7 or 9 × 9. For the example shown in Fig. 1, the vector representation of pixel x_5 is $\boldsymbol{X} = (x_1, x_2, ..., x_9)^T$.

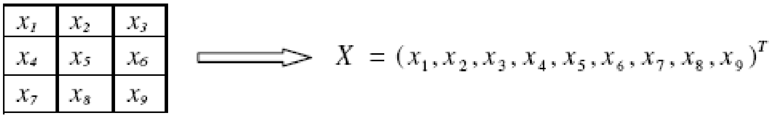


Fig. 1. A central pixel and its neighbors are illustrated on the left and its substituted vector is illustrated on the right.

More generally, let x_i denote the intensities of the pixels in the sliding window of the reference image, and y_i denote those of the pixels in the corresponding window of the current image. Then the vector representations are $\boldsymbol{X} = \{x_i, i = 1, 2, ..., n\}$ and $\boldsymbol{Y} = \{y_i, i = 1, 2, ..., n\}$ respectively. As indicated by the shading model, if there is no change, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linear dependent. Accordingly, it is easy to prove that the variance σ^2 in the sliding window expressed as below should be zero.

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{y_i}{x_i} - \mu\right]^2, \quad \mu = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}. \tag{4}$$

Following this work, many other test criteria for linear dependence were also proposed. For instance, in [7] the criterion is designed on top of the determinants of Wronskian matrices (we call it Wronskian detector in the following discussions). If there is no change, no matter with or without illumination, this test should be some constant $k_0(k_0 - 1)$ (see Section 2.3 for the definition of k_0).

$$W = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{x_i^2} - \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}. \tag{5}$$

2.3 Noise Analysis

Although the above linear dependence based methods are expected to give illumination invariant change detection results, this is only true when no noise is present. In practical situations, the digital images are usually interfered by the noises of cameras and reflections. In such cases, as shown below, these algorithms will encounter problems.

Here we use a popular noise model, in which the noise has additive Gaussian distribution. When there is no change, the image intensities can be written as:

$$\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{\delta_1}, \quad \boldsymbol{Y} = k_0\boldsymbol{S} + \boldsymbol{\delta_2}, \tag{6}$$

where $\boldsymbol{S}$ denotes the vector of the underlying image which is not affected by noise, and k_0 is a scalar factor representing the linear dependence. $\boldsymbol{\delta_1}, \boldsymbol{\delta_2}$ are two noise vectors in which the distribution of each element is $N(0, \sigma_d^2)$(i.i.d). By working out $\boldsymbol{S}$ from the first equation in (6) and substituting it to the second one, we can get the following model:

$$\boldsymbol{Y} = k_0\boldsymbol{X} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = \boldsymbol{\delta_2} - k_0\boldsymbol{\delta_1}, \tag{7}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ is the combined noise vector in which the distribution of each element is $N(0, \sigma_s^2)$ (i.i.d). Substitute the first equation of (7) to (4), there flows:

$$\mu^* = k_0 + \frac{1}{n}\sum_{i=1}^{n}\frac{\epsilon_i}{x_i}; \tag{8}$$

$$(\sigma^*)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big[\frac{y_i}{x_i} - \mu^*\Big]^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n}\frac{\epsilon_i^2}{x_i^2} - \frac{1}{n}\Big(\sum_{i=1}^{n}\frac{\epsilon_i}{x_i}\Big)^2\Big]. \tag{9}$$

As the elements of $\boldsymbol{\epsilon}$ are independently identically distributed, there holds:

$$E[\epsilon_i\epsilon_j] = \begin{cases} E[\epsilon_i]E[\epsilon_j] = 0, \; i \neq j \\ E[\epsilon_i^2] = \sigma_s^2, \qquad i = j \end{cases} \tag{10}$$

Hence, the expectation of $(\sigma^*)^2$ is:

$$E[(\sigma^*)^2] = \frac{1}{n-1}\Big[\sum_{i=1}^{n}\frac{E[\epsilon_i^2]}{x_i^2} - \frac{1}{n}E[\Big(\sum_{i=1}^{n}\frac{\epsilon_i}{x_i}\Big)^2]\Big] = \frac{\sigma_s^2}{n}\sum_{i=1}^{n}\frac{1}{x_i^2}. \tag{11}$$

Similar noise analysis can be done to the criterion of the Wronskian detector (WD) and its expectation is:

$$E[W^*] = k_0(k_0 - 1) + \frac{\sigma_s^2}{n}\sum_{i=1}^{n}\frac{1}{x_i^2}. \tag{12}$$

From the above derivations, we could find that even if there is no change, the expectations of the test criteria used in LDD and WD vary along with the image attributes and the noise variances. As a result, it is difficult to select a reasonable threshold. Most likely, the selection can only be done empirically and adaptive to different images. To tackle this problem, an unbiased criterion is required. In order to do that, we propose a new method in the next section, where the technique of linear approximation is adopted.

3 Linear Approximation Detector (LAD)

3.1 The Proposed Method

For the purpose of applying linear approximation, an assistant plane is used with an orthogonal coordinate system. The components of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are reorganized as points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ in the assistant plane. Considering that the above n points might coincide with each other, an extra helper point $(x_0, y_0) = (0, 0)$ is added to the points set. The linear approximation of these points is calculated using least square algorithm [9], the corresponding result line of which is as below,

$$y = kx + b; \tag{13}$$

$$k = \Big[(n+1)\sum_{i=0}^{n} x_i y_i - \sum_{i=0}^{n} x_i \sum_{i=0}^{n} y_i\Big] \Big/ \Big[(n+1)\sum_{i=0}^{n} x_i^2 - \Big(\sum_{i=0}^{n} x_i\Big)^2\Big]; \tag{14}$$

$$b = \frac{1}{n+1}\Big[\sum_{i=0}^{n} y_i - k\sum_{i=0}^{n} x_i\Big]. \tag{15}$$

Following the idea of shading model, if there is no change, vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ should be linear dependent. That is, we could use b as a test criterion: there is no change when $|b| = 0$; otherwise, a change is detected. We call the proposed method as linear approximation detector (LAD).

3.2 Noise Analysis

When the same noise exists as described in section **2.3**, we could get the following derivations (the details are removed due to limitation of the paper length).

$$k^* = k_0 + (n+1)\sum_{i=0}^{n} x_i \epsilon_i \Big/ \Big[(n+1)\sum_{i=0}^{n} x_i^2 - \Big(\sum_{i=0}^{n} x_i\Big)^2\Big]; \tag{16}$$

$$b^* = \frac{1}{n+1}\Big[\sum_{i=0}^{n}(k_0 x_i + \epsilon_i) - k^*\sum_{i=0}^{n} x_i\Big] \tag{17}$$

$$= \sum_{i=1}^{n} x_i \epsilon_i \Big[\sum_{i=1}^{n} x_i \Big/ \Big((n+1)\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2\Big)\Big]. \tag{18}$$

We could further get the conclusion that the expectation of the criterion b^* is zero and it can be hardly affected by both the noise and the local pixel intensities.

$$E[b^*] = \sum_{i=1}^{n} x_i E[\epsilon_i] \Big[\sum_{i=1}^{n} x_i \Big/ \Big((n+1)\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2\Big)\Big] = 0. \tag{19}$$

In this sense, the proposed LAD algorithm will be more robust to noise than the aforementioned LDD and WD methods. This does make sense because in fact the noise always exists. And for some practical cases, especially when the environment is dark, the influence of the noise is much more significant.

4 Experimental Results

In our experiments, the algorithms of LDD, WD and LAD were implemented on the frames of *PETS2001*[1]. The size of the original color images in CAMERA1 is 768 × 576 pixels. We convert them to gray images and resize them to 384 × 288 pixels. For the experimental settings, we used a fixed sliding window of size 7×7.

Firstly, we tested the algorithms' performance on detecting changes between real images. Fig. 2(a) is the reference image (FRAME 0001), and Fig. 2(b) is the current image (FRAME 1220, where the car in the street corner moves back a little and an extra car moves in). Fig. 3(a) to (c) are the detection results by LDD, WD and LAD respectively. From these results, we can see that LDD and WD lead to some false alarms, while LAD can well wipe off these influences. However, the objects detected by LAD are a bit smaller than their actual sizes due to the operation of linear approximation.

Secondly, we would like to test the illumination sensitivity of the above algorithms. Fig. 2(c) is gained by adding illumination artificially to Fig. 2(b) with the scalar factor k_0 increasing gradually from the rightmost column ($k_0 = 0.9$) to the leftmost column ($k_0 = 1.3$). The change detection results between Fig. 2(a) and Fig. 2(c) are shown in Fig. 3(d) to (f). From them, we can see that all of these three algorithms are illumination-invariant. This is just the design purpose of utilizing linear algebra.

Thirdly, we examined the detectors' performance with camera noises added. For this purpose, we generate Fig. 2(d) by adding zero-mean Gauss noises to Fig. 2(b). Fig. 3(g) to (i) show that the noises are sensitively detected as moving objects by LDD and WD, while LAD successfully wipes off the influence of the noises.

Fourthly, we enumerated different thresholds for LDD, WD and LAD algorithms and got the *recall-precision* curves by comparing the detection results to the manually labelled ground truth in Fig. 4(a). The *recall* and *precision* are calculated as:

$$recall = D/(D+M), \quad precision = D/(D+F), \tag{20}$$

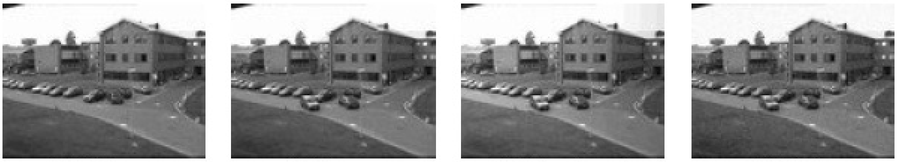
where D, M, and F denote the pixel numbers of the accurate detections, the missed detections and the false alarms respectively. Fig.4.(b) tells us that LAD outperforms LDD and WD by much in most of the cases. For example, when the precision is 80%, LAD resulted in 50% higher recall than LDD and WD.

Lastly, we fixed the thresholds for each algorithm and got the $F1$-*noise* curves by changing the variance of the added Gaussian noises from 0 to 10, where $F1$ is defined as:

$$F1 = 2 \times (recall \times precision)/(recall + precision). \tag{21}$$

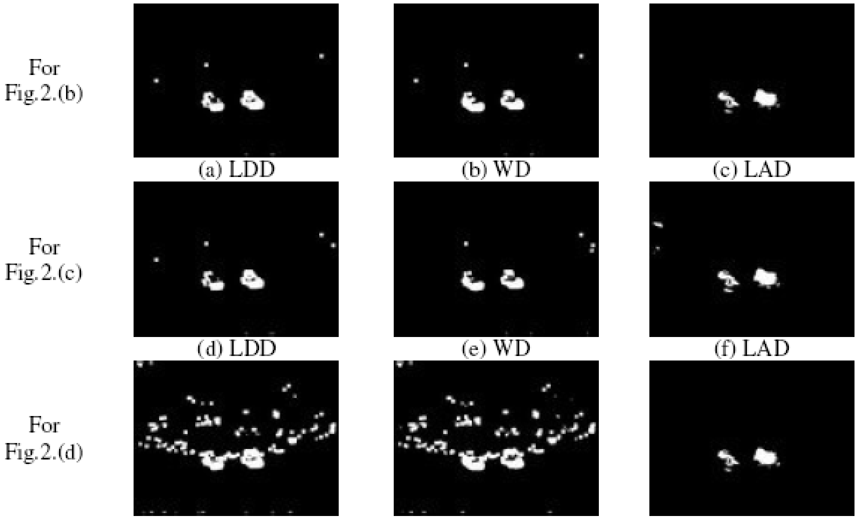
We can see from Fig. 4(c) that the $F1$ curves of LDD and WD drop quickly when the noises become heavier. However, the curve of LAD is not influenced much by the noises, indicating to some extend that LAD is quite robust to noises.

[1] http://peipa.essex.ac.uk/ipa/pix/pets/PETS2001/DATASET1/TRAINING/



(a) Reference (b) Current (c) Illuminated (d) Noisy

Fig. 2. Test images used in the experiments.

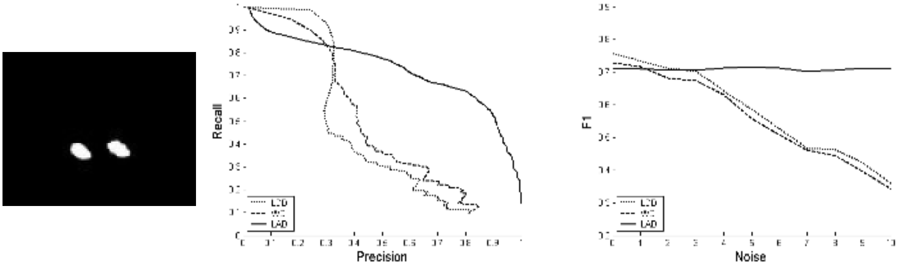


(a) LDD (b) WD (c) LAD

(d) LDD (e) WD (f) LAD

(g) LDD (h) WD (i) LAD

Fig. 3. Detection results of LDD, WD and LAD algorithms (here we manually adjusted the threshold to get best compromises between *recall* and *precision*).



(a)Ground truth (b)*recall-precision* curves (c)$F1$-*noise* curves

Fig. 4. The ground truth, *recall-precision* and $F1$-*noise* curves for LDD, WD and LAD.

5 Conclusion

This paper introduces a new linear algebra based approach for image change detection by adopting the linear approximation technique. All linear dependence based methods in the paper are analyzed theoretically with a popular noise model, which shows that the proposed method can avoid the intrinsic weakness of the other congener methods. Experimental results also verify that the proposed algorithm is much more robust to camera noises and reflections than other linear algebra based methods.

References

[1] T. Aach, A. Kaup: Statistical model-based change detection in moving video. Signal Processing, vol. 31, pp. 165–180, 1993

[2] T. Aach, A. Kaup, R. Mester: Change detection in image sequences using Gibbs random fields. Proc. IEEE, Workshop on Intelligent Signal Processing and Communication Systems, Sendai, Japan, pp. 56–61, Oct. 1993

[3] T. Aach, A. Kaup: Bayesian algorithms for adaptive change detection in image sequences using Markov random fields. Signal Processing: Image Communication, vol. 7, pp. 147–160, 1995

[4] Y. Z. Hsu, H. H. Nagel, G. Reckers: New likelihood test methods for change detection in image sequences. Computer Vision, Graphics, and Image Processing: vol. 26, pp. 73–106, 1984

[5] K. Skifstad, R. Jain: Illumination independent change detection for real world image sequences. Computer Vision, Graphics, and Image Processing: vol. 46, pp. 387–399, 1989

[6] E. Durucan, T. Ebrahimi: Robust and illumination invariant change detection based on linear dependence. Proc. of 10th European Signal Processing Conference, Tampere, Finland, pp. 1141-1144, Sept. 2000

[7] E. Durucan, T. Ebrahimi: Change detection and background extraction by linear algebra. Proc. IEEE 89(10): pp. 1368–1361, Oct. 2001

[8] B. T. Phong: Illumination for computer generated pictures. Commun. ACM, Vol. 18, pp. 311-317, 1975

[9] Richard L. Burden, J. Douglas Faires: Numerical Analysis. Brooks Cole, 2000

3D Model Similarity Measurement with Geometric Feature Map Based on Phase-Encoded Range Image

Donghui Wang[1] and Chenyang Cui[2]

[1] Dept. of Computer Science, Zhejiang University
Hangzhou, Zhejiang, P.R.China, 310027
mii@cs.zju.edu.cn

[2] National Key Lab. Of CAD&CG, Zhejiang University
Hangzhou, Zhejiang, P.R.China, 310027
ccy@cad.zju.edu.cn

Abstract. Measuring the similarity between 3D models is a very important problem in 3D model retrieval. A challenge aspect of this problem is to find a suitable shape descriptor that can grasp the feature of 3D model. In this paper, Geometric Feature Map of 3D model based on phase-encoded range-image is proposed. This map contains the information about the surface normal of the model and the area proportion of the planar surface of the models .From the local map of model at every possible rotation, we can obtain a global map of the model. The similarity calculation between 3D models is processed using a coarse-to-fine strategy,the similarity calculation is fast and efficient. The experimentation result shows that our method is invariant to translation , rotation and scaling of the model and agree with general human intuition, and particularly useful for classification of 3D models.

1 Introduction and Related Work

Recent development in modelling and digitizing techniques has led to an increasing accumulation of 3d models. So, the problem of matching 3D models has become the recent focus of research efforts. The feature extraction of 3D model is the fundamental problem in similarity matching. There are many techniques for shape matching and feature extraction. Mahmoudi et al. (see [1,2,3 and 4]) use curvature distribution as the feature of 3D models, some good results have been provided for uses such as 3D shape pose estimation, however, the curvature is sensitive to noise and small undulation on the model surface and the choice of the number of views(see [1]) which characterize a 3d model is a problem too. Novotni M et al.(see [5]) propose a specific distance histograms that define a measure of geometric similarity of the inspected objects, this technique is limited to the specific applications, and inadequate for a general 3D shape search. Hilaga et al. (see [6]) use reeb graph based on geodesic distance which represents the topology of the model for 3D model matching, however, the reeb graph doesn't cover full geometric information and the topology of the graph is

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 103–110, 2004.

dependent on the choice of interval size, simultaneously, the geodesic distance is unsuitable for all 3D models. Osada et al.(see [7]) propose a continuous probability distribution as shape signature for 3D model based on the shape functions such as the angle between 3 random points on the surface , the advantage of this method is that no complex feature extraction is necessary, and robust to small perturbation on the model boundary. However, such distributions do not grasp the intrinsic features of the model since it use the surface feature of the model instead of the volume feature of the model. Ohbuchi et al. (see [8]) use a combination of three vectors (such as the moment of inertia , the average distance of the surface from the axis and the variance of distance of the surface from the axis) as the feature of the model, the experiment results show that this method is more suitable for the symmetric models . This paper proposes a method called geometric feature map based on phase-encoded range-image of 3D models. It is invariant to the translation and rotation of 3D model. Principal component analysis (PCA) is firstly used to find a best view line, from which the Initial Position Model (IPM) is obtained (see Figure 1), then resampling the IPM to the range image. we encode the range image as phase, a planar surface of 3D model will become a linear phase factor, then, the Geometric Feature Map (GFM) of the model is obtained by the Fourier transform of the phase-encoded range-image. For every possible rotation of the model based on the IPM, the corresponding GFM can be obtained, so, a unique feature map of the model is defined. Experimentation result shows our method is suitable for coarse-to-fine matching of 3D models.

2 Feature Extraction

3D models are usually given in arbitrary units of measurement and in unpredictable position and orientations in 3D-space. In order to find the best matching of two models, three steps are needed: (1). Find the best view line and reduce the consume of 3D model retrieval. This step ensures that the initial range image from the given best view line is obtained. The Modified Principal Component Analysis is firstly applied to change the coordinate system axes to the new ones which are consistent with the three large spreads of the vertex distribution, to find the best view line which defines the z axis. (2). Fourier transform of phase-encoded range-image. This step ensures that the extracted feature is invariant to the translation of the model. The range image of the model is firstly encoded as phase, after phase encoding, Fourier transform of the range image is only relevant to the orientation of the normal of the planar surface. (3). Fourier transforms under limited rotation of models. This step ensures that the extracted feature is invariant to the rotation of the model. The IPM from the given best view line is continuously rotated around a few regular orientations, then the corresponding phase-encoded range-images are obtained, the corresponding Fourier transforms are calculated.

2.1 Find the Best View Line

Let $P = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_i, ..., \boldsymbol{p}_n\}$ be the set of the vertices of the given polygon model, here, $\boldsymbol{p}_i \in R^3, i = 1, 2, ..., n$, and n is the number of vertices. A symmetric and real covariance matrix M is defined as:

$$M = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{p}_i * \boldsymbol{p}_i^T \tag{1}$$

Here, $\boldsymbol{p}_i^T$ is the transpose of $\boldsymbol{p}_i$. Find the three eigenvalues of matrix M and sort them in decreasing, then a transform matrix M_t is got by three feature vectors corresponding to three eigenvalues. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. That means, the first feature vector corresponding to the maximum eigenvalue represents the orientation of x axis in the new coordinate frame, the second feature vector corresponding to the middle eigenvalue represents the orientation of y axis, and the third feature vector corresponding to the minimum eigenvalue represents the orientation of z axis. Here, z axis is defined as the best view line. Finally, we apply the matrix M_t to transform the original position and orientation of 3D model to the new ones $\boldsymbol{p}_{ni}$:

$$\boldsymbol{p}_{ni} = M_t * \boldsymbol{p}_i \tag{2}$$

See Figure 1, the left plane is the original model before PCA, the right is the IPM from the best view line after PCA. In Figure 1, z axis is defined as the best view line. In 3D model retrieval, the IPM of the given model is firstly used as the searching key, in order to get better searching results, a coarse-to-fine strategy is used.

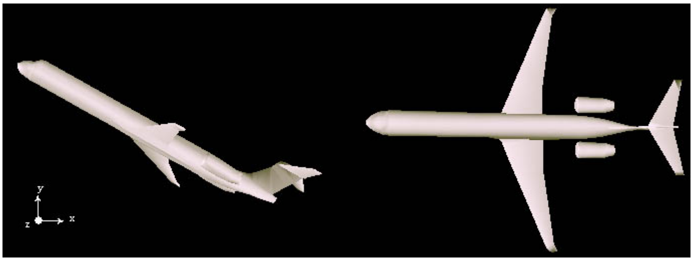


Fig. 1. The original model and the Initial Position Model

2.2 Fourier Transforms of Phase-Encoded Range Image

When resampling a model to the range image $z = f(x, y)$,some parts of the model may be hidden in the image,here,only the points whose z-value is bigger than zero are preserved. The range image contains the depth information of the model. Firstly, we encode the range image as phase as the following equation:

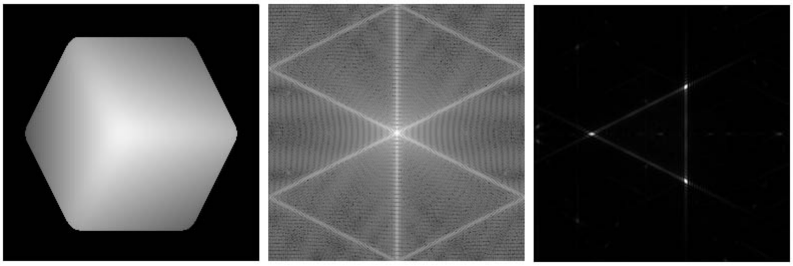


Fig. 2. the range image of the cube and the Fourier transform and the Fourier transform of phase-encoded range image

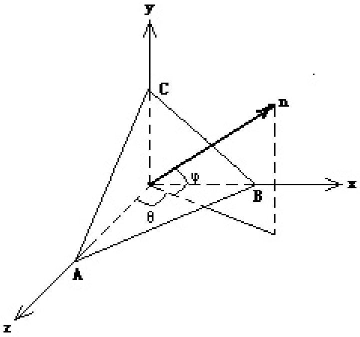


Fig. 3. the definition of the normal of the surface

$$ph(x_k, y_k) = exp(iwf(x_k, y_k)), k = 1, 2, ..., n \tag{3}$$

Here,n is the maximum sampling frequency, w is a parameter that adjusts the phase slope of the model. In this paper we assume $w = 1$. The Fourier transform of the phase-encoded range image(phFT) $phFT(u_k, v_k)$ is:

$$phFT(u_k, v_k) = F_{2D}(exp(if(x_k, y_k))) \tag{4}$$

F_{2D} represents two dimensional Fourier transform. See Figure 2, the left is the range image of a cube , where, only three surfaces are visible, the middle is the Fourier transform of the range image,the right is the phFT of the range image, where, the three peaks represents three visible surfaces of the cube. As shown in Figure 3, the orientation of the normal n of the planar surface ABC can be defined by two angles (θ, ϕ),so, we can describe (u_k, v_k) as the following:

$$(u_k, v_k) = (\frac{\tan(\phi_k)}{2\pi}, \frac{\tan(\theta_k)}{2\pi\cos(\phi_k)}) \tag{5}$$

Finally, we have:

$$phFT(u_k, v_k) = phFT(\theta_k, \phi_k) \tag{6}$$

From Figure 2 and equation (6), we know the Fourier transform of the phase-encoded range image is dependent on the orientation of the normal of the planar surface of the model, and independent of the position of the model.

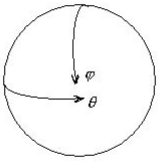


Fig. 4. the orientation of the rotation of the model

2.3 Fourier Transforms Under Limited Rotation

If the full 3D views of the model from arbitrary view line are known, we can calculate the phFT for any possible orientation. This makes it possible to obtain a global feature information about the normals for all possible angles in a single image by displacing and pasting the phFT expressed in spherical coordinates. 3D model matching will be invariant to not only the translation of the model but also the rotation of the model. In order to obtain the full information of the model under arbitrary orientation, we continuously rotate the IPM around the two directions (see Figure 4): one is around ϕ direction at 30 degree steps (ϕ is from 45 degree to 135 degree), the other is around the θ direction at 30 degree steps(θ is from 0 degree to 180 degree). The new model after rotating Q is described as the following:

$$Q = \{\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_i, ..., \boldsymbol{q}_n\}, i = 1, 2, ..., n \tag{7}$$

$\boldsymbol{q}_i$ is the new vertex of the model point $\boldsymbol{p}_i$ after rotating:

$$\boldsymbol{q}_i = R(\phi) \cdot R(\theta) \cdot \boldsymbol{p}_i \tag{8}$$

where, $R(\phi)$ and $R(\theta)$ are the rotation matrices around ϕ and θ orientation respectively.

$$R(\phi) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\phi) & -\sin(\phi) & 0 \\ 0 & \sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9}$$

$$R(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

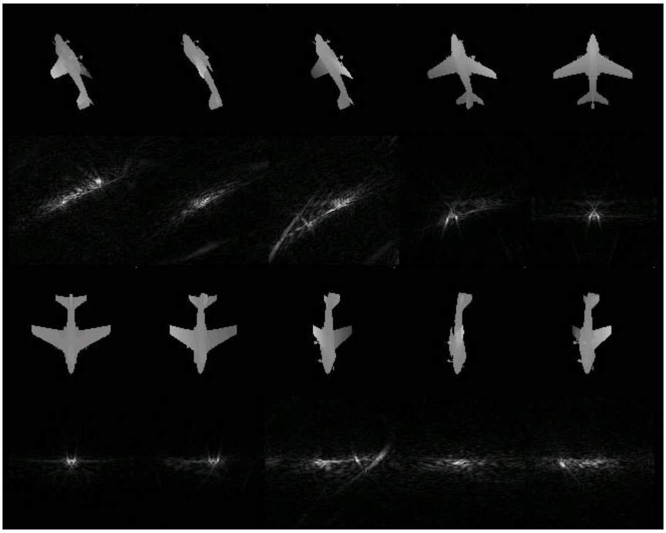


Fig. 5. The range images and the corresponding phFTs from the different view lines

After rotating, different orientation range images are obtained. Figure 5 shows the part of the range images of the model at all possible rotational orientation and the corresponding phFTs. This assures that the information of the model under any view line is obtained.

3 Similarity Measurement

Similarity among a pair of models is computed as the distance between their corresponding phFTs. The Euclidean distance is very common distance function for high-dimensional feature vectors, where the individual components of the feature vectors are assumed to be independent from each other, and no relationships of the components maybe regarded. In our algorithm, we emphasize the relationship between the two components, and avoid the shortcoming of the classic Euclidean Distance. The modified Euclidean distance d between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ can be computed by using the following equation:

$$d_A^2(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})\cdot(A)\cdot(\boldsymbol{x}-\boldsymbol{y})^T = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(\boldsymbol{x}_i-\boldsymbol{y}_i)(\boldsymbol{x}_j-\boldsymbol{y}_j) \qquad (11)$$

here, A is a similarity matrix where the components a_{ij} of the matrix A represent the similarity of the components i and j in the underlying vector space ,and a_{ij} can be calculated by using the formula $a_{ij} = \exp^{-\sigma\cdot d(i,j)}$, the parameter σ controls the global shape of the similarity matrix. In our experiments, the distance D between any two models is:

$$D = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{12}$$

where, d_i represents the modified Euclidean distance between the corresponding phFT from the same view line for every pair of compared models, which can be computed by equation (11). The similarity between two 3D models $Score\_Similarity$ ($score\_Similarity \in [0,1]$)is described as the following:

$$Score\_Similarity = 1 - D \tag{13}$$

4 Experiment Results

For the experiment, we used 220 3DS models which were collected from the Internet. In this case, in order to reduce the computation cost, the two steps are needed: 1. The phFTs of the IPMs for each of the 220 models are obtained in advance. One model is selected as the search key from the 220 models, the similarities between the search key and the other remaining models are calculated, and the models are sorted according to the resulting similarity. 2. The phFTs at all rotational orientation of the resulting models from the first step and the search model are calculated. Some example results of the experiment are shown in Figure 6. The selected model is shown as the key and the models returned with the highest similarities are shown under searched models. 220 models are classified into 30 categories.The experiment results show that the method based on phFT can accurately identify models, especially not be influenced by the rotation ,translation, scaling and simplification of the model.The similarity matching agrees with general human intuition.

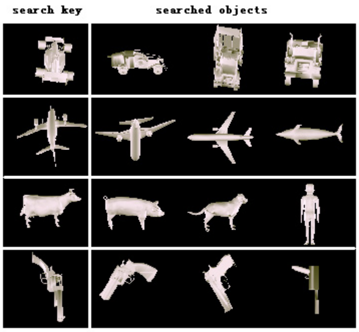


Fig. 6. Results of the search experiment

5 Discussion

In this paper, we presented a new technique called Geometric Feature Map based on phase-encoded range image. The experiments indicate that our method provide a fast and efficient computation of the similarity between models and provides results that agree with human intuition. Currently, the similarity measurement between two models are calculated one by one according to the corresponding a series of range images. In the future, we expect the spherical Fourier transforms will be used to extract the features of the models, which can overcome the shortcoming of 2D Fourier transform and improve the accuracy of the extracted features from the models and the speed of the similarity matching between two models.

References

1. Mahmoudi,S., Daoudi M.: 3D Models retrieval by using characteristic Views. IEEE Computer Graphics & Applications. (2002) 1051–4651.
2. Sonthi,R., Kunjur G., Gadh R.: Shape feature determination using the curvature region representation. Proc. Symp. Solid Modeling. (1997) 285–296.
3. Kang S., Ikeuchi K.: The complex EGI:new representation for 3-D pose determination. IEEE Trans. PAMI, Vol.15. (1993) 707–721.
4. Wang W., Iyengar S.: Efficient data structures for model-based 3-D object recognition and localization from range images. IEEE Trans. PAMI, Vol.14. (1992) 1035–1045.
5. Novotni M., Klein R.: Geometric 3D comparison and application. Proceedings of ECDL WS Generalized Documents. (2001) 39–44.
6. Hilaga M., Shinagawa U., Kohmura T., Kunii T.L.: Topology matching for fully automatic similarity estimation of 3D shapes. Proc. SIGGRAPH, (2001)
7. Osada R., Funkhouser T., Chazelle B., Dobkin D.: Matching 3D models with shape distributions. International Conference on Shape Modeling and Applications. (2001)
8. Ohbuchi R., Otagiri T., Ibato M.: Shape-similarity search of three-dimensional models using parameterized statistics. Proceedings of the Pacific graphics. (2002) 265–273.
9. Garcia J., Valles J., Ferreira G.: Detection of three-dimensional objects under arbitrary rotations based on range images. OPTICS EXPRESS Vol.11, No.25,(2003).
10. http://www.fon.hum.uva.nl/praat/manual/Principal_component_analysis.html

Automatic Peak Number Detection in Image Symmetry Analysis*

Jingrui He[1], Mingjing Li[2], Hong-Jiang Zhang[2],
Hanghang Tong[1], and Changshui Zhang[3]

[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, hjzhang}@microsoft.com
[3] Automation Department, Tsinghua University, Beijing 100084, P.R.China
zcs@tsinghua.edu.cn

Abstract. In repeated pattern analysis, peak number detection in autocorrelation is of key importance, which subsequently determines the correctness of the constructed lattice. Previous work inevitably needs users to select peak number manually, which limits its generalization to applications in large image database. The main contribution of this paper is to propose an optimization-based approach for automatic peak number detection, i.e., we first formulate it as an optimization problem by a straightforward yet effective criterion function, and then resort to Simulated Annealing to optimize it. Based on this approach, we design a new feature to depict image symmetry property which can be automatically extracted for repeated pattern retrieval. Experimental results demonstrate the effectiveness of the optimization approach and the superiority of symmetry feature over wavelet feature in discriminating similar repeated patterns.

1 Introduction

Repeated pattern symmetry has been studied for decades and plays a nontrivial role in texture analysis. In a 2D repeated pattern, there exists a finite region bounded by two linearly independent vectors. Repeating along those two vectors, it produces simultaneously a covering (no gaps) and a packing (no overlaps) of the original image [5]. The finite region is the repeated unit and the two vectors are called translation vectors, which build up lattice structure. According to the theory of wallpaper groups, the infinite variety of repeated patterns can be well categorized into seventeen Crystallographic groups, which are characterized by four kinds of symmetry: translation symmetry, rotation symmetry, reflection symmetry and glide reflection symmetry [1]. Among them, rotation symmetry can only be 2-fold, 3-fold, 4-fold and 6-fold, where n-fold means $360/n$ degree rotation. Reflection and glide reflection symmetry can have four axes: two translation vectors and two diagonal vectors of the repeated unit. The fundamental

* This work was performed at Microsoft Research Asia.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 111–118, 2004.

problem in repeated pattern analysis is to determine whether or not it has a certain kind of symmetry, which relies on the construction of translation vectors, and to further classify it into one of the seventeen groups.

In lattice construction, Leung et al. [3] and Schaffalitzky et al. [6] both use parameterized affine transforms to establish correspondence between interesting regions. Starovoitov et al. [7] propose a more traditional approach which uses features extracted from co-occurrence matrix to detect translation vectors. It has been proved in [4] that if a texture image is of regular structure, autocorrelation is more appropriate than Fourier Transform in analyzing its structure. More recently, Liu et al. [5] first calculate the autocorrelation of the pattern, and then select a sparse set of points to designate the repeated units. Finally, they adapt a Hough transform approach [4] to find two translation vectors from the points. To select candidate points from autocorrelation, they propose an efficient approach based on the region of dominance instead of simple threshold method and the points are called peaks. For isolated patterns, this method proves to be most appropriate [5]. However, the problem of automatic peak number detection remains unsolved. The drawback limits its generalization to applications in large image database, as it tends to be a tedious task to manually set peak number for each image in the database. Although peak number is apparent to people, determining its value automatically is difficult due to the variability of autocorrelation of different patterns.

In this paper, we propose an optimization-based scheme to automatically select peak number. To be specific, firstly, we formulate peak number selection as an optimization problem by a straightforward yet effective criterion, which incorporates the highest score in the accumulator array, autocorrelation value, as well as the length of the translation vectors. Then, we resort to Simulated Annealing (SA) to optimize it in order to balance between optimization performance and processing time. Based on this scheme, we design a new symmetry feature using translation vectors for repeated pattern retrieval which can be applied in spin industry where users wish to find repeated patterns with similar symmetry property as the query from a large database. Experimental results demonstrate the effectiveness of our method.

The rest of this paper is organized as follows. In Section 2, we present our approach of automatic peak number detection; in Section 3, the symmetry feature extraction is proposed; experimental results are given in Section 4; finally, we conclude the paper in Section 5.

2 Automatic Peak Number Detection

To construct translation vectors correctly, peak number N should be an approximation of the number of repeated units in an image. Once peak number is determined, the Generalized Hough Transform (GHT) [4] can be utilized to find the two translation vectors [5]. The procedure can be summarized as follows. Initially, a 2D accumulator array is created, in which each entry is set to be zero. Secondly, each pair of non-collinear vectors are used as two translation vectors to span a parallelogram grid. For each peak, if it is located near any vertex of the parallelogram grid, the score that the peak belongs to the parallelogram grid will be high; otherwise the score will be low. Thirdly, the score is added to the entry corresponding to the pair of non-collinear vectors. Finally, two translation vectors are obtained by finding the entry with the highest score in the accumulator array.

In order to automatically detect peak number, we design an optimization-based approach. In this approach, the peak number is obtained by optimizing a criterion function. Details are presented below.

2.1 Criterion Function

Let the entry with the highest score in the accumulator array S_N locates at (i, j), where N denotes the peak number. When N is proper, almost all of the selected peaks are located at the lattice node, thus $S_N(i, j)$ will be fairly large. However, as N increases, more peaks will be selected, and more scores will be added to the entries of S_N. So the original maximum corresponding to proper peak number N will be overwhelmed. Thus we divide $S_N(i, j)$ by N to eliminate the accumulation effect and take this value as a criterion for selecting proper peak number N.

However, in GHT, the accumulating score is inverse proportional to the lengths of two translation vectors. In patterns with sub-units, translation vectors of sub-units are always shorter than those of real units. If peak number is well above proper, the criterion value will be larger than that of proper peak number and sub-units will be extracted. Inspired by the fact that the autocorrelation value of peaks corresponding to sub-units is not as conspicuous as that of real units, we add a height factor to the criterion to help distinguish real units from sub-units.

When peak number is small, one of the translation vectors may be zero, which is unreasonable. So we add another term $\min(|w_1|, |w_2|)$ to the criterion, which selects the shorter one of the two translation vectors to preclude unreasonable pair of vectors.

Based on the above discussion, the final criterion function can be written as follows:

$$C(N) = (\frac{S_N(i, j)}{N})^{\alpha} \cdot (\overline{height(N)})^{\beta} \cdot (\min(|w_1|, |w_2|))^{\gamma} \tag{1}$$

where $\overline{height(N)}$ is the average autocorrelation value of the first N peaks, i.e. the height factor; α, β, γ are positive parameters controlling the contribution of the three terms to the overall criterion. In our current implementation, they are set to 1 for simplicity.

2.2 Optimization of the Criterion Function

As the criterion function contains several local maxima (Fig. 2(d)), traditional greedy algorithms may fail to find the global maximum. Although enumeration

```
1. Initialize peak number $N = N_0$, $temperature = t_0$, $time = 0$. Calculate $C(N_0)$, and set maximum criterion value $C_{max} = C(N_0)$, optimum peak number $N_{opt} = N_0$;
2. Repeat $n_1$ times:
   (a) Set $N1 = N + \Delta N$, where $\Delta N$ is an integer distributed uniformly distributed uniformly between $[-\Delta, +\Delta]$. Calculate $C(N_1)$. If $C(N_1)$ is bigger than $C(N)$, set $N = N_1$; otherwise, set $N = N_1$ with probability $\exp((C(N_1) - C(N))/temperature)$;
   (b) Update $C_{max}$ and $N_{opt}$;
3. Decrease $temperature$ by $\Delta t$. If maximum criterion value $C_{max}$ has not changed for $n_2$ times, output $N_{opt}$ that produces $C_{max}$ and stop; else go to 2.
```

Fig. 1. SA algorithm for peak number detection

methods can always find the optimal solution, it is very time-consuming due to the construction of accumulation array. To balance optimization performance and processing time, we use Simulation Annealing (SA) algorithm, which is listed in Fig. 1.

3 Symmetry Feature Extraction

To determine whether or not repeated patterns have a certain kind of symmetry, Liu et al. [5] apply the symmetry to be tested to the entire pattern, and check the similarity between the original and transformed images. However, repeated patterns are often corrupted by noise or distortion. Therefore, it is more reasonable to give a continuous value to measure the extent to which a pattern has a certain kind of symmetry than a yes or no conclusion.

3.1 Symmetry Measure

If a pattern has a certain kind of symmetry, the correlation between the original and transformed images will have peaks with similar underlying structure as autocorrelation. To make this structure similarity concrete, we extract translation vectors from both autocorrelation and correlation, and compare the two pairs of vectors.

Let w_1, w_2 denote the translation vectors calculated from autocorrelation. t_1, t_2 denote the translation vectors calculated from correlation between the original and transformed images. Using automatic peak number detection method discussed in Section 2, we can automatically determine how many peaks should be selected from autocorrelation and correlation, and construct translation vectors without the intervention of users. If a pattern has a certain kind of symmetry, the associated t_1 and t_2 will be approximately the same as w_1 and w_2, or sometimes a rotated version; otherwise they will be totally different. To correctly measure the similarity between the two pairs of translation vectors, we use three components to represent each pair: the lengths of w_1 and w_2 (t_1 and t_2) and the angle between them. Then the chessboard distance is calculated to measure the similarity:

$$D(w_1, w_2, t_1, t_2) = \frac{1}{|w_1|}||w_1| - |t_1|| + \frac{1}{|w_2|}||w_2| - |t_2|| + |\theta_w - \theta_t| \qquad (2)$$

where θ_w (θ_t) is the angle between w_1 and w_2 (t_1 and t_2) in radian. When calculating the distance automatically, we do not know if w_1 corresponds to t_1 or t_2, so we calculate both $D(w_1, w_2, t_1, t_2)$ and $D(w_1, w_2, t_2, t_1)$, and select the smaller the one as the final distance. We also normalize the distance to [0,1] by formalizing the following exponential form:

$$S = \exp(-\min(D(w_1, w_2, t_1, t_2), D(w_1, w_2, t_2, t_1))) \qquad (3)$$

If an image has a certain kind of symmetry, S will be near 1; otherwise, it will be small, sometimes near 0. For an image not strictly symmetrical due to noise or distortion or some other reason, S measures to what extent this pattern has a certain kind of symmetry.

3.2 Symmetry Feature

Every repeated pattern has translation symmetry. The translation vectors constructed from autocorrelation surface reflect this symmetry. To measure rotation symmetry likelihood, we perform the four kinds of rotation to the original image, construct translation vectors based on correlation, and get four measures S_2, S_3, S_4, S_6 using equation (3), where s_n denotes the symmetry measure for n-fold rotation symmetry. Reflection symmetry and glide reflection symmetry are essentially the same except for a translation factor. So we perform reflection transformations using each choice of the axes, construct translation vectors and get another four measures S_{T1}, S_{T2}, S_{D1}, S_{D2}, where footnote T denotes translation vector axes, and D denotes diagonal vector axes. The eight measures make up for the symmetry feature, which represents the symmetrical property of a repeated pattern, and can be written as follows:

$$f_S = [S_2, S_3, S_4, S_6, S_{T1}, S_{T2}, S_{D1}, S_{D2}]^t \qquad (4)$$

4 Experimental Results

4.1 Peak Number Detection

In our experiment, we have collected 487 repeated patterns from the web. These images include all seventeen wallpaper groups. The number of images belonging to each group is listed in Table 1. From the table, we can see that the number varies greatly from group to group, which reflects the non-uniformity of various kinds of symmetry in natural images. Most of the patterns are corrupted by noise or distortion, thus not strictly symmetrical. Figure 2 illustrates two examples of automatic peak number detection. Note that: 1) in both cases, the maximum

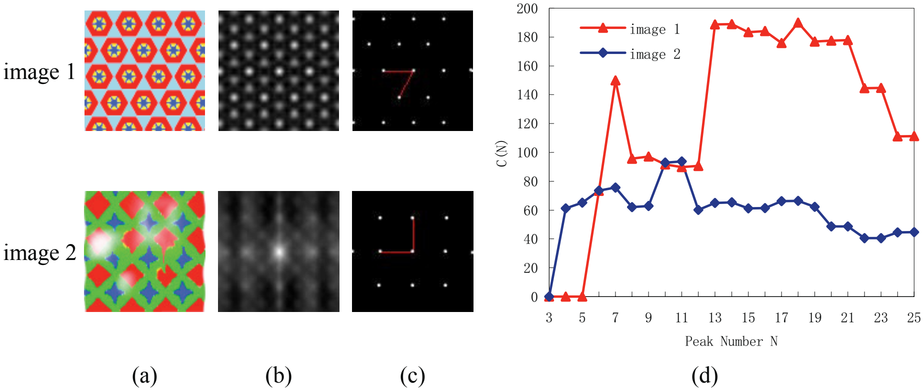


Fig. 2. Examples of automatic peak number detection via optimizing the criterion function. (a) repeated patterns [2]; (b) autocorrelation; (c) N_{opt} peaks and the translation vectors; (d) criterion value $C(N)$ (due to limited space, we only exhibit its value with peak number N from 3 to 25; when N is larger than 25, $C(N)$ is small and lacks variability)

value of the criterion is achieved at the proper peak number; 2) the response in the neighborhood of the proper peak number is somewhat flat, however it is still applicable for real applications (for explanation, see the next paragraph); 3) in both cases, there exists local maximum points, which indicates a greedy algorithm might fail to find the optimal number.

In order to test the performance of the criterion function, we first use an enumeration method to search for the optimal peak number that corresponds to maximum criterion value and extract the translation vectors. In 437 patterns, the constructed translation vectors are consistent with human perception, which means that the selected peak number is correct. Accordingly, the correct rate is 89.73%. For comparison, we also test the performance of the SA algorithm running several times. All the parameters are determined based on some prior knowledge of the database. In our experiment, we set $N_0 = 13$, $\Delta N = 5$, $t_0 = 0.01$, $\Delta t = 0.0001$, $n_1 = 3$, $n_2 = 5$. Currently we are doing research on optimal selection of these parameters. On average, 436 pairs of translation vectors are correctly extracted, i.e. the correct rate is 89.53%, which is very close to that of the enumeration approach. The number of patterns whose translation vectors are wrongly detected for each group by the two algorithms are also listed in Table 1. It is worth noticing that the image whose translation vectors are wrongly detected by enumeration method is NOT necessarily wrongly detected by SA method. This can be explained as follows: the robustness of GHT ensures that if peak number N lies in the neighborhood of repeated unit number N_0, i.e. $N \in [N_0 - \Delta N_1, N_0 + \Delta N_2]$, the constructed translation vectors will be correct. (This also explains the aforementioned problem) If peak number N_{opt} corresponding to maximum criterion value does not lie in this range, translation vectors will be unstable in case of peak number perturbation. Since SA is a kind of random search strategy, it may often return peak number in $[N_0 - \Delta N_1, N_0 + \Delta N_2]$ and

Table 1. Optimization method comparison. E: Enumeration, S: SA

Wallpaper groups		p1	p2	pm	pg	cm	pmm	pmg	pgg	cmm	p4	p4m	p4g	p3	p31m	p3m1	p6	p6m
Number		22	25	11	5	40	46	13	15	49	33	117	9	6	11	13	12	60
Error Num	E	1	10	1	2	3	3	2	0	5	2	8	0	0	2	1	1	9
	S	2	7	1	0	6	3	1	0	6	2	13	0	1	2	1	0	6

construct correct translation vectors. On the other hand, the overall correct rate of SA is less than that of enumeration. This may partially due to the prematurity property of SA.

As mentioned before, the enumeration scheme is very time consuming. For an image containing 72 × 128 pixels, the time needed with peak number varying from 3 to 100 is about 43 seconds (Intel(R) 500MHz, 256M RAM). But the optimization time of SA for the same image is less than a second on average.

4.2 Symmetry Feature Evaluation

To test the performance of symmetry feature in repeated pattern retrieval, we build this feature for the 487 repeated patterns, use each of the patterns as a query, and record average precision. The repeated patterns are classified into seventeen wallpaper groups as ground truth. We use wavelet feature for comparison, which is a widely accepted descriptor of texture. It consists of 18 coefficient moments, which is obtained after three-level Daubechies-8 wavelet transforms [8]. Both results are illustrated in Table 2. The precision values are not very high. That is because most of the repeated patterns are corrupted by noise or distortion. Therefore, although peak number is correctly identified, the underlying structure of correlation between the original and transformed images does not resemble that of autocorrelation, thus t_1 and t_2 will not be correctly calculated. For this reason, we only focus on the relative performance in subsequent discussion.

Comparing symmetry feature and wavelet feature, although the latter is a widely accepted descriptor of texture, and has more dimensions, its performance is not as good as symmetry feature. Moreover, the first retrieved images using symmetry feature are visually more similar with the query than those using wavelet feature. It demonstrates the effectiveness of our feature from another point of view.

Table 2. Feature performance comparison

Precision	P10	P20	P30	P40	P50
Symmetry Feature	0.1840	0.1738	0.1684	0.1647	0.1605
Wavelet Feature	0.1777	0.1567	0.1462	0.1416	0.1368

5 Conclusion

In this paper, we have proposed an optimization-based scheme for automatic peak number detection, which enables repeated patterns to be analyzed without human intervention. Moreover, we design a new symmetry feature which reflects the sym-metrical property of repeated patterns. Based on automatic peak number detection method, this feature can be extracted automatically for a large number of images. Currently we are doing research on generalizing this symmetry feature to natural images.

Acknowledgements. This work was supported by National High Technology Research and Development Program of China (863 Program) under contract No.2001AA114190.

References

[1] Gruenbaum, B., Grunbaum, B., Shephard, G.C.: Tilings and Patterns. W.H. Freeman and Company, New York (1987)

[2] Kali: Programs that can automatically generate 2D planar crystallographic patterns. http://www.geom.umn.edu/apps/kali/

[3] Leung, L., and Malik, J.: Detecting, localizing and grouping repeated scene elements. ECCV LNCS 1064 **1** (1996) 546–555

[4] Lin, H., Wang, L., and Yang, S.: Extracting periodicity of a regular texture based on autocorrelation functions. Pattern Recognition Letters **18** (1997) 433–443

[5] Liu, Y., and Collins, R.: A computational model for repeated pattern perception using Frieze and Wallpaper groups. Proc. CVPR **1** (2000) 537–544

[6] Schaffalitzky, F. and Zisserman, A.: Geometric grouping of repeated elements within images. Proc. BMVC (1998) 13–22

[7] Starovoitov, V.V., Jeong, S.Y., and Park, R.H.: Texture periodicity detection: Features, properties, and comparisons. IEEE SMC-A **28(6)** (1998) 839–848

[8] Unser, M.: Texture classification and segmentation using wavelet frames. IEEE Trans. on Image Processing **4** (1995) 1549–1560

Image Matching Based on Singular Value Decomposition

Feng Zhao[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences,
Beijing 100080, China
[2] Graduate School of the Chinese Academy of Sciences,
Beijing 100039, China
fzhao@jdl.ac.cn

Abstract. This paper presents a simple and effective method for matching two uncalibrated images. First, corner points are extracted in both images separately. Then the initial set of point matches is obtained by singular value decomposition of a well-designed correspondence strength matrix. A new expression of this matrix is introduced to get more reliable initial matches. Last, the epipolar geometry constraint is imposed to reject the false matches. Experimental results on real images show this method to be effective for general image matching.

1 Introduction

Matching two uncalibrated images is a classic problem in computer vision. The topic has been researched for several decades. One effective strategy is using interest point matching technology. The interest point matching technology first extracts points of interest such as corners in the two images and then uses robust matching technology to establish point correspondences between the two images [1,2,3,4].

Corners are considered as good candidates for interest points in many computer vision applications such as motion correspondence, object tracking and stereo matching, etc. Using corner points as interest points has been proved to be effective in image matching [1,2]. Among the most popular corner detectors, Harris corner detector [5] is known to be robust against rotation and illumination changes. And its results have high repeatability under different imaging conditions, which is of great advantage to image matching task [6].

In the recent literature some effective strategies for interest point correspondence between image pairs have emerged. A robust technique for matching two uncalibrated images has been proposed by Zhang et al. [1], which finds initial matches using correlation and relaxation methods followed by the LMedS technique to discard false matches. Pilu [2] used a direct method based on singular value decomposition for feature correspondence, which was originally proposed by Scott and Longuet-Higgins [7]. The method first sets up a correlation-weighted proximity matrix and then performs singular value decomposition calculation on

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 119–126, 2004.

the matrix to get initial matches. We improve Pilu's method by giving a new expression of the matrix. Experiments on real images show the improved method can handle more complicated cases and obtain more reliable initial matches.

The initial set of interest point matches usually contains some false matches due to the bad locations of corners and the improper matches in the establishment of correspondence. Constraints like epipolar geometry or disparity gradient limit can be used to reject the false matches [1,8,9].

This paper presents a simple and effective method to solve the problem of image matching. The method is based on singular value decomposition. The initial set of point matches is directly obtained by singular value decomposition of a well-designed correspondence strength matrix. A new expression of the correspondence strength matrix is introduced. The comparison of matching results on real images demonstrates the new matrix outperforms that used in previous works. Experimental results on real images show the method is effective for matching image pairs under different imaging conditions.

The remainder of the paper is organized as follows. Section 2 describes extracting corners as interest points with slightly modified Harris corner detector. Section 3 introduces the matching approach based on singular value decomposition and presents the new expression of the correspondence strength matrix. Section 4 describes rejecting the false matches by imposing epipolar geometry constraint. Section 5 presents some experimental results on real images and Section 6 concludes the paper.

2 Extracting Corners as Interest Points

Corners are highly informative image locations. They are very useful features for many computer vision applications. Corners can be automatically detected without any prior knowledge and they are stable features for image matching task. Many algorithms for detecting corners have been reported up to now. Among the most popular corner detectors, Harris corner detector [5] is known to be robust against rotation and illumination changes. The results of Harris corner detector have high repeatability under different imaging conditions, which is very important for image matching.

Harris corner detector is based on the auto-correlation matrix, which is built as follows:

$$M = \exp -\frac{x^2+y^2}{2\sigma^2} \otimes \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} , \tag{1}$$

where $\otimes$ is the convolution operation. I_x and I_y indicate the x and y directional derivative respectively. The auto-correlation matrix performs a smoothing operation on the products of the first derivatives by convolving with a Gaussian window. Two sufficient large eigenvalues of M indicate the presence of a corner. To avoid the explicit eigenvalues decomposition of M, a corner response function is defined by the following expression:

$$C_H = \det(M) - k * \mathrm{trace}(M)^2 , \tag{2}$$

where k is usually set to 0.04. Then a threshold t_h can be used to select corner points. If $C_H > t_h$ then the point is identified as a corner.

In the original Harris detector, I_x and I_y are computed by convolution with the mask [-1 0 1] and the corner response function requires setting a parameter k. In our implementation, the mask [-2 -1 0 1 2] is used to compute the first derivatives. And a corner measure function without additional parameter [10] is adopted. The measure function is defined as follows:

$$C_F = \frac{\det(M)}{\mathrm{trace}(M)} \ . \tag{3}$$

A threshold t_f is given and a point is considered as a corner if $C_F > t_f$. In our implementation, t_f is set to 1% of the maximum observed C_F.

3 Matching Based on Singular Value Decomposition

The problem of matching interest points in two images is fundamental in computer vision. A direct method for establishing feature correspondences between two images based on singular value decomposition (SVD) has been proposed in Scott and Longuet-Higgins [7] and further developed in Pilu [2]. The basic idea of this method is obtaining feature correspondences with singular value decomposition of a correspondence strength matrix. In the following, the algorithm will be briefly described.

Let I_1 and I_2 be two images. I_1 contains m features $I_{1i}(i = 1 \ldots m)$ and I_2 contains n features $I_{2j}(j = 1 \ldots n)$. To get one-to-one feature correspondences between the two images, a proximity matrix $G \in M_{m,n}$ is set up first:

$$G_{ij} = \ e^{-r_{ij}^2/2\sigma^2} \ , \tag{4}$$

where $r_{ij} = \parallel I_{1i} - I_{2j} \parallel$ is the Euclidean distance between the two features if they are regarded as being on the same plane. G is a positive definite matrix and the element of G decreases monotonically from 1 to 0 with the increase of the distance between the two features. The degree of interaction between the two sets of features is controlled by the parameter σ.

The next step of the algorithm is to perform the singular value decomposition of G:

$$G = \ UDV^T \ , \tag{5}$$

where $U \in M_{m,m}$ and $V \in M_{n,n}$ are orthogonal matrices and the diagonal matrix $D \in M_{m,n}$ contains the singular values along its diagonal elements D_{ii} in descending order.

Then a new matrix E is converted from the diagonal matrix D by replacing every diagonal element of D with 1. Computing the following product will get the new matrix H:

$$H = \ UEV^T \ . \tag{6}$$

The matrix $H \in M_{m,n}$ has the same shape as the proximity matrix G. Two features I_{1i} and I_{2j} are pairing up if H_{ij} is both the greatest element in its row and the greatest element in its column. With selecting all such elements in H, the feature correspondences between the two images can be established.

Scott and Longuet-Higgins algorithm only takes spatial location into account for establishing feature correspondences. To include similarity information in feature matching with above algorithm, Pilu [2] uses normalized cross-correlation score as local measurement to quantify feature similarity. Adding similarity constraint can eliminate rogue features, which shouldn't be similar to anything. The elements of G can then be transformed as follows:

$$G_{ij} = \frac{(C_{ij}+1)}{2}\, e^{-r_{ij}^2/2\sigma^2} , \tag{7}$$

where C_{ij} is the normalized cross-correlation score between the two features.

In our method, a new expression of the correspondence strength matrix G is introduced in order to get more reliable initial matches under different imaging conditions. The new definition of the correspondence strength matrix G is given as follows:

$$G_{ij} = (C_{ij}+1)^3\, e^{-r_{ij}/2\sigma^2} , \tag{8}$$

where σ is set to 50 in our system. C_{ij} can be calculated as follows:

Let $m_1 = I_1(x_i, y_i)$ be the i-th interest point in the first image I_1 and $m_2 = I_2(x_j, y_j)$ be the j-th interest point in the second image I_2. W_1 and W_2 are two windows of size $(2w+1)\times(2w+1)$ centered on each point. The normalized cross-correlation score C_{ij} is defined as:

$$C_{ij} = \frac{\sum_{u=-w}^{w}\sum_{v=-w}^{w}\left[I_1(x_i+u, y_i+v) - \overline{I_1(x_i,y_i)}\right]\left[I_2(x_j+u, y_j+v) - \overline{I_2(x_j,y_j)}\right]}{(2w+1)(2w+1)\sigma(m_1)\sigma(m_2)} , \tag{9}$$

where $\overline{I_1(x_i,y_i)}$ ($\overline{I_2(x_j,y_j)}$) is the average and $\sigma(m_1)$ ($\sigma(m_2)$) is the standard deviation of all the pixels in the correlation window. The size of the correlation window is 11 in our system.

The new definition of G gives more weight to the similarity measurement (i.e. normalized cross-correlation score) and weakens the spatial location measurement. Experiments on complicated real images show the new expression of G yields better results. The initial set of matches contains less false matches than the former methods and the number of good matches increases simultaneously.

4 Rejecting False Matches

The initial set of interest point matches usually contains some false matches. Constraints like epipolar geometry can be used to reject the false matches. The epipolar geometry constraint can be described as follows.

Suppose F is the fundamental matrix. Point $p\,(x, y)$ in the image can be represented as:

$$\widetilde{p} = [x, y, 1]^T . \tag{10}$$

For a point match (p, q), the epipolar line of point p in the first image is defined as:

$$l_p = F\widetilde{p}\,. \tag{11}$$

If the match is perfect, then point q in the second image should lie on the epipolar line l_p exactly. The distance d_q of point q to the epipolar line l_p is calculated by

$$d_q = \frac{|\ \widetilde{q}^T F\widetilde{p}\ |}{\sqrt{(F\widetilde{p})_1^2 + (F\widetilde{p})_2^2}}\ , \tag{12}$$

where $(F\widetilde{p})_i$ is the i-th component of vector $F\widetilde{p}$. The distance d_p of point p to the epipolar line l_q is calculated similarly. Then a threshold t_e can be used to find the good matches. A point match is identified as a good match if $\max(d_p\,, d_q) \le t_e$.

In our implementation, the epipolar geometry constraint is imposed based on RANSAC [11]. The method is described as follows. First, eight matches are randomly chosen from the initial set of point matches and the fundamental matrix F is calculated from them. Then we can find all good matches that are consistent with the epipolar geometry constraint according to this fundamental matrix. The threshold t_e is 1.5 in our system. Repeat the above steps to get the largest set of good matches.

5 Experimental Results

Experiments on real images of various content have been performed. Some of the results are reported here. These images are under different imaging conditions such as rotation, scale changes, illumination changes and the combination of them.

All the results here are obtained with the same parameter setting as mentioned in the forgoing sections. These results are the final results after applying RANSAC on the initial matches. Image pairs **Boat**, **Residence**, **Car** are from

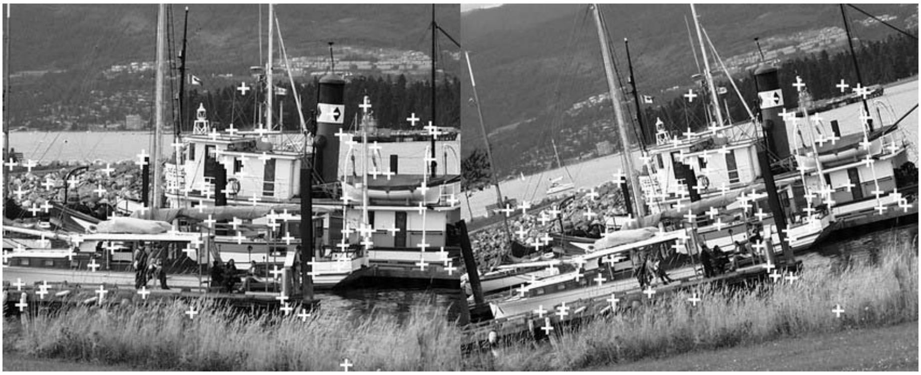


Fig. 1. Matching result for image pair **Boat** with rotation and scale changes

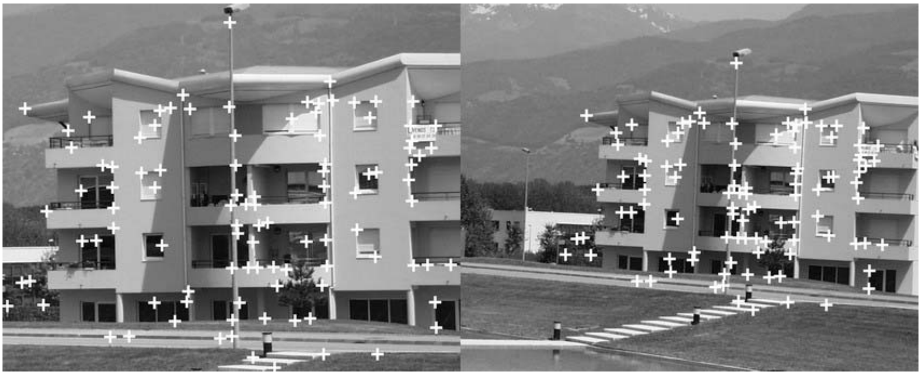


Fig. 2. Matching result for image pair **Residence** with rotation and scale changes

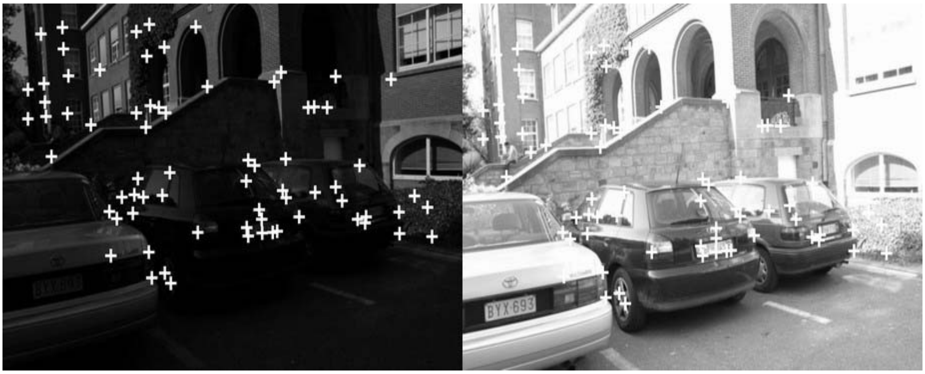


Fig. 3. Matching result for image pair **Car** with illumination changes

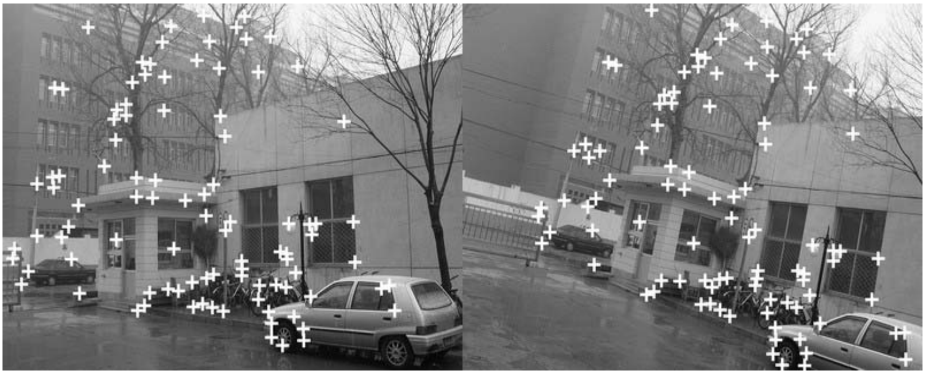


Fig. 4. Matching result for image pair **Gate** with translation and rotation

Table 1. The columns of "Final Matches" contain the number of good matches after rejecting the false matches in the initial matches with epipolar geometry constraints. In addition, the algorithm using original G fails in matching the image pair **Gate**, therefore the corresponding number of initial matches and final matches are small

	Original G		**Improved G**	
	Initial Matches	**Final Matches**	**Initial Matches**	**Final Matches**
Boat	88	65	109	91
Residence	89	51	133	102
Car	85	55	100	72
Gate	45	10	133	96

INRIA[1], and image pair **Gate** is from our lab. More reliable matching results have been obtained by using the new expression of the correspondence strength matrix G . Table 1 gives the comparison of matching results between original definition of G (Equation (7)) and the improved G (Equation (8)).

6 Conclusions

We have presented a simple and effective method for matching two uncalibrated images. The method exploits singular value decomposition, which is one of the stablest numerical matrix operations. The kernel of this method is establishing interest point correspondences by singular value decomposition of a well-designed correspondence strength matrix. We introduce a new expression of this matrix to get more reliable initial matches. The comparison of matching results on real images demonstrates the new matrix outperforms that used in previous works. One characteristic of this algorithm is that it does not require a correlation threshold for a candidate point match to be accepted. This contributes to the simplicity of the method. Experimental results on real images show the method is effective for matching image pairs under different imaging conditions.

Acknowledgements. This work is supported by National Hi-Tech Development Programs of China under grant No. 2003AA142140.

References

1. Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry. *Artificial Intelligence Journal*, Vol. 78, pp. 87–119, 1995
2. M. Pilu. A direct method for stereo correspondence based on singular value decomposition. *IEEE CVPR*, pp. 261–266, 1997

[1] http://lear.inrialpes.fr/people/Mikolajczyk/Database/index.html

3. A. Baumberg. Reliable feature matching across widely separated views. *IEEE CVPR*, pp. 774–781, 2000
4. D. Tell and S. Carlsson. Combining Appearance and Topology for Wide Baseline Matching. *7th European Conference on Computer Vision (ECCV)*, pp. 68–81, 2002
5. C. Harris and M. Stephens. A combined corner and edge detector. *Fourth Alvey Vision Conference*, pp. 148–151, 1988
6. C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, Vol. 37, pp. 151–172, 2000
7. G. Scott and H. Longuet-Higgins. An algorithm for associating the features of two patterns. *Proceedings of the Royal Statistical Society of London*, Vol. B244, pp. 21–26, 1991
8. G. Xu, E. Nishimura, and S. Tsuji. Image correspondence and segmentation by epipolar lines: Theory, algorithm and applications. Technical report, Dept. of Systems Engineering, Osaka University, Japan, 1993
9. S.B. Pollard, J.E.W. Mayhew, and J.P. Frisby. PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, Vol. 14, pp. 449–470, 1985
10. W. Faustner. A feature based correspondence algorithm for image matching. *International Archives of Photogrammetry and Remote Sensing*, Vol. 26, pp. 150–166, 1986
11. M. A. Fischler and R. C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, Vol. 24, pp. 381–395, 1981

Image Matching Based on Scale Invariant Regions

Lei Qin[1], Wei Zeng[2], and Weiqiang Wang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] Department of Computer Science and Technology, Harbin Institute of Technology, China
{lqin,wzen,wqwang}@jdl.ac.cn

Abstract. In this paper we present a novel approach to match two images in presenting large scale and rotation changes. The proposed approach is based on scale invariant region description. Scale invariant region is detected by a two-step process and represented by a new descriptor. The descriptor is a two-dimensional gray-level histogram. Different descriptors can be directly compared. In addition, our descriptor is invariant to image rotation and large scale changes as well as robust to small viewpoint changes. The experiments demonstrate that our method is effective enough for image matching and retrieval.

1 Introduction

Image matching, which is a fundamental aspect of many computer vision problems, has been extensively studied for the last two decades. The image matching problem estimates correspondences between two views of the same scenes, where the viewpoints differ by position, orientation and viewing parameters. It is a challenging task due to the various appearances that an image can have. Many global approaches have been proposed in the literatures to solve this problem [3,12]. But they have difficulty in dealing with nearby clutter and partial occlusion. Recently local information has been shown to be sufficient to describe image content [1], and well adapted to image matching, as they are robust to clutter and occlusion and don't require segmentation [2,5,6,7].

Schmid and Mohr [1] demonstrate that local information is sufficient for general image recognition under occlusion and clutter. They use Harris corners [10] to detect interesting points, and extract a rotationally invariant descriptor from the local image region. This guarantees features to be matched between rotated images. Lowe [2,16] uses the local extrema of difference-of-Gaussian in scale-space as the interesting points. He proposes a distinctive local descriptor, which is computed by accumulating local image gradients in orientation histograms. Tuytelaars and Van Gool [7] construct small affine invariant regions around corners and intensity extrema. Their method looks for a specific structure "parallelogram" in images. Among the above methods, [1] and [2] are rotation and scale invariant. [7] is affine invariant.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 127–134, 2004.

Many methods have been presented for wide-baseline matching [6,8,9,11]. Baumberg [9] uses the multi-scale Harris features detector, and orders the features based on the scale-normalized feature strength. The neighborhood of feature point is normalized using an iterative procedure based on isotropy of the second gradient moment matrix. Mikolajczyk and Schmid [8] propose the Harris-Laplace method. They first detect the Harris corners in multiple scales, then select points at which the Laplacian measure attained the local extrema in scale dimension. They extend their work to affine invariant in [5]. Schaffalitzky and Zisserman [6] present a method for obtaining multi-view matching given unordered image sets. They use two kinds of features: invariant neighbourhoods and "Maximally Stable Extremal" regions. They use the complex filters as descriptor. J. Xiao, and M. Shah [11] combine Canny edge detector and Harris corner operator together to present a new edge-corner interesting point. Based on SVD decomposition of affine matrix, a two-stage matching algorithm is proposed to compensate the affine transformation between two frames. They use the epipolar geometry to estimate more correspondences interesting points.

In this paper we want to solve the problem of matching images in the presence of both large scale and rotation changes. Mikolajczyk and Schmid try to solve this problem in [8]. They use Harris-Laplace detector to select the position and scale of interesting points, and extract rotation invariant descriptors. Since they use multi-scale space to select interesting points, their method is suitable for significant scaling case. But before computing the distance of descriptors, they must learn a distance metric in invariant space from training data. So the metric is tuning to the domain of training data. The matching results depend on the efficiency of learned metric. In order to overcome this problem, we propose a novel descriptor, SC descriptor, based on Shape Context [17]. The new descriptor is scale and rotation invariant. The distance between SC descriptors can be calculated directly.

The paper is organized as follows. In Section 2, we introduce the Harris-Laplace detector. In Section 3, we present a novel descriptor based on Shape Context [17]. In Section 4, we describe the robust matching algorithm. In Section 5, we demonstrate the effectiveness of our approach for image matching and retrieval. In section 6 we summarize the contribution of this paper and draw some conclusions.

2 Harris-Laplace Detector

The task of image matching by local information requires detecting image regions, which are co-variant with scale transformations of the image. A two-step algorithm is used to achieve the co-variant regions (x, y, $scale$): 1) detecting interesting points (x, y), 2) associating a characteristic scale to each interesting point ($scale$).

2.1 Interesting Points Detector

Computing image descriptors for each pixel in image is time consuming and unnecessary. The descriptor extraction is limited to a subset of the image pixels, interesting points, which should have two main properties: distinctiveness and invariance. Different interesting points detectors are comparatively evaluated in [15]. The Harris corner detector is the most reliable one in the presence of image rotation, illumination changes, and perspective transformations. The basic idea of this detector is to use the auto-correlation function in order to determine locations where the signal changes in two directions. A matrix related to the auto-correlation function which takes into account the first derivatives of the signal on a window is computed. The eigenvectors of this matrix are the principal curvatures of the auto-correlation function. Two significant values indicated the presence of an interesting point.

However the repeatability of this detector degrades significantly when the images have large scale changes. In order to cope with such changes, a multi-scale Harris detector is presented in [13].

Multi-scale Harris function $det(C) - \alpha trace^2(C)$

With $C(\mathtt{x}, \sigma_I, \sigma_D) = \sigma_D^2 G(\sigma_I) * \begin{pmatrix} L_x^2(\mathtt{x}, \sigma_D) & L_x L_y(\mathtt{x}, \sigma_D) \\ L_x L_y(\mathtt{x}, \sigma_D) & L_y^2(\mathtt{x}, \sigma_D) \end{pmatrix}$,
where σ_I is the integration scale, σ_D the derivation scale, $L(\mathtt{x}, \sigma) = G(\sigma) * I(\mathtt{x})$ different levels of resolution created by convoluting the Gaussian kernel $G(\sigma)$ with the image $I(\mathtt{x})$ and $\mathtt{x} = (x, y)$. Given an image $I(\mathtt{x})$ the derivatives can be defined by $L_x(\mathtt{x}, \sigma) = \frac{\partial}{\partial x} G(\sigma) * I(\mathtt{x})$. We use the multi-scale Harris detector to detect interesting points at different resolution $L(\mathtt{x}, \sigma_n)$, with $\sigma_n = k^n \sigma_0$, $\sigma_0$is the initial scale factor.

2.2 Scale Selection

Lindeberg [14] suggests that for Geometric image descriptors, the characteristic scale can be defined as that at which the result of a differential operator is maximised. The characteristic scale is relatively independent of the image scale. The ratio of the scales, at which the extrema were found for corresponding points in two rescaled images, is equal to the scale factor between the images. Different differential operators are comparatively evaluated in [8]. Laplacian obtains the highest percentage of correct scale detection. We use the Laplacian to verify for each of the candidate points found on different levels if it forms a local maximum in the scale direction.

$Lap(\mathtt{x}, \sigma_n) > Lap(\mathtt{x}, \sigma_{n-1}) \wedge Lap(\mathtt{x}, \sigma_n) > Lap(\mathtt{x}, \sigma_{n+1})$
Figure 1 shows two images with detected interesting points. The threshold of multi-scale Harris and Laplacian are 1000 and 10, respectively. 15 resolution levels are used for scale representation. The factor k is 1.2 and $\sigma_I = 2\sigma_D$.

3 Invariant Descriptor

Given an image region which is co-variant with scale transformations of the image, we wish to compute a descriptor which is both invariant to scale and

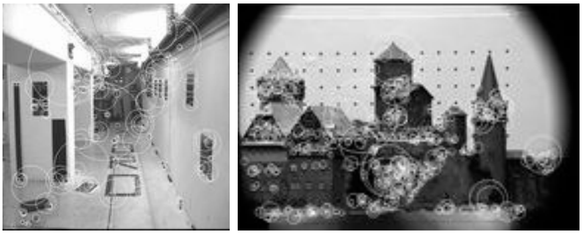


Fig. 1. Scale invariant regions found on two images. There are 252 and 417 points detected in the left and right images, respectively

rotation changes and distinctive to reduce the ambiguity in correspondences. The descriptor we use is novel, and we now discuss this.

3.1 SC Descriptor

In this paper, we propose a novel invariant region descriptor, the SC descriptor, inspired by the work of Shape Context [17] introduced by Belongie, Malik and Puzicha for matching shapes. A SC descriptor is a two-dimensional histogram encoding the intensity distribution in an image region, and the geometry relation between the sub-regions. The two dimensions of the histogram are d, the sequence number of n sub-regions, and i, the intensity value. The sequence of sub-regions is sorted in ascending order both in and as shown in Fig. 2(a). Each column of the SC descriptor is the intensity histogram of the corresponding sub-region. We use bins that are not uniform in , which are selected by experiments. An example is shown in Fig. 2. To achieve invariance to affine transformations of the intensity (transformations of the form $I \rightarrow aI + b$), it is sufficient to normalize the range of the intensity within the sub region.

Our region detector is a local approach and does not need segmentation, so it is robust to clutter and occlusion. For the same reason the invariance to translation is intrinsic to our approach. Due to multiple window sizes and selecting

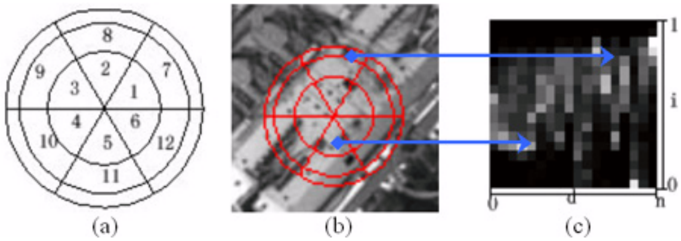
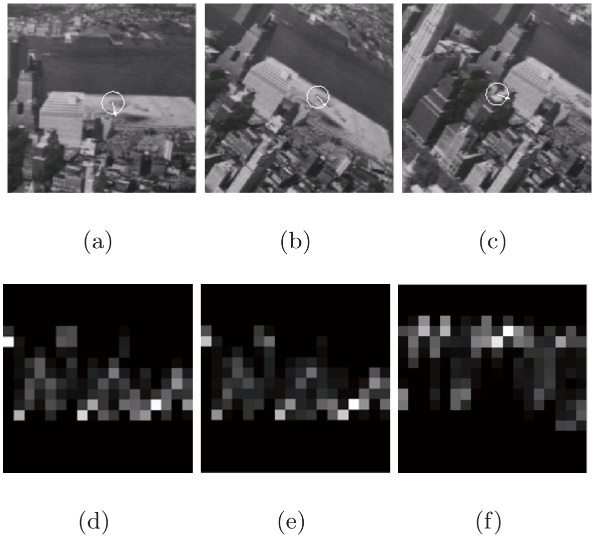


Fig. 2. SC descriptor construction. (a) A image region is divided into 18 sub-regions. We use three bins for $\rho(0.57, 0.85, 1)$ and six bins for $\theta(1/3\pi, 2/3\pi, \pi, 4/3\pi, 5/3\pi, 2\pi)$. The number in the sub-region is its sequence number. The sequence numbers of the third ring aren't shown for clarity. Two sub-regions in an image patch (b) map to two different columns in the SC descriptor (c)



(a) (b) (c)

(d) (e) (f)

Fig. 3. More examples of SC descriptor. Top: image patch (white circle) and the approximate gradient direction (white arrow). (b) is a rotation version of (a). (b) and (c) are different parts of the same image. Bottom: (d), (e) and (f) are SC descriptors of (a), (b) and (c), respectively. SC descriptor is a two-dimensional histogram. Each column of SC descriptor is a histogram of a bin of Fig.2(a). (White =large value). Note that (d) and (e) are similar, which are calculated for relatively similar region except rotation. While (f) is quite different with (d) and (e). We don't normalize the intensity here

scale automatically, scale invariance is achieved. The invariance to rotation can be obtained by normalizing the region with respect to the gradient direction, or by using a relative frame, which treats the gradient direction as the positive x-axis. One can use the method in [4] to get a stable estimation of the dominant direction. However we have observed that the SC descriptor is robust to small error in gradient direction estimation, as the gradient directions in Fig. 3(a) and Fig. 3(b) are different, but the SC descriptors are similar. We use the gradient direction $\theta_g = arctan(L_y/L_x)$. After turning the relative frame with the gradient phase, the descriptor is completely rotation invariant. The number of sub-region should be chosen according to specific application. Figure 3 shows several examples.

3.2 Comparison of SC Descriptor

Let P_i and Q_j represent two image regions. And $COST_{ij} = COST(P_i, Q_j)$ denotes the cost of matching these two regions. We use χ^2 test statistics [17].

$$COST_{ij} \equiv COST(P_i, Q_j) = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=0}^{255} \frac{[h_i(n,m) - h_j(n,m)]^2}{h_i(n,m) + h_j(n,m)} \qquad (1)$$

where $h_i(n,m)$ and $h_j(n,m)$ denote the value of the SC descriptor at P_i and Q_j, respectively.

4 Robust Matching

To robust match a pair of images, we first determine point-to-point correspondence. We select for each descriptor in the first image the most similar one in the second image based on the cost function (1). If the cost is below a threshold the correspondence is kept. All point-to-point correspondences form a set of initial matches. We refine the initial matches using RANdom SAmple Consensus (RANSAC). RANSAC has the advantage that it is largely insensitive to outliers. We use fundamental matrix as the transformation of RANSAC in our experiments.

5 Experimental Results

5.1 Robust Image Matching

The matching strategy just described is applied and tested over a large number of image pairs. All of our experiments use gray images to compute correspondences. And we give the epipolar geometries of two images estimated by our method, to show that our matching algorithm is effective.

Figure 4 shows the matching results for three different scenes, which include significant rotation and scale changes. Figure 4(a) and Fig. 4(c) show scale changes between images. Figure 4(b) shows both rotation and scaling between two images. The rotation angle of Fig. 4(b) is 10 degrees, and scale factor is 2.4. Figure 5 displays two images with estimated epipolar lines. All the epipolar geometry is correctly estimated.

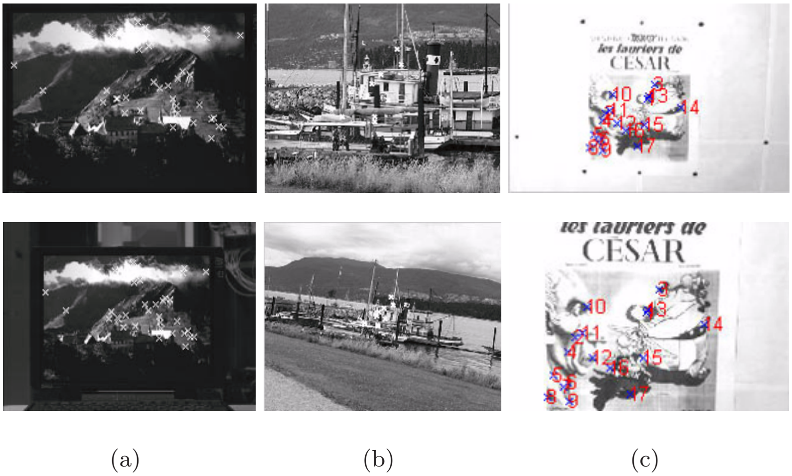


Fig. 4. Robust matching results by our algorithm. (a)(frames of "Laptop" from INRIA) shows 34 inliers, 32 of them are correct. (b) (frames of "boat" from INRIA) shows 9 inliers, all of them are correct. (c) (frames of "Asterix" from INRIA) shows 17 inliers, all of them are correct

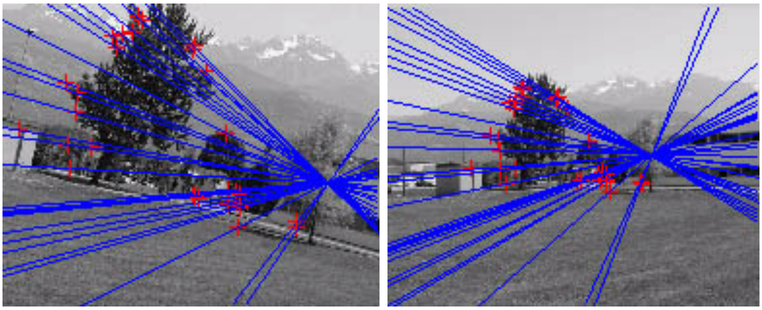


Fig. 5. This figure shows the epipolar geometry as computed with the matching method described in this paper. Two frames of ”East” from INRIA show 30 correct inliers.

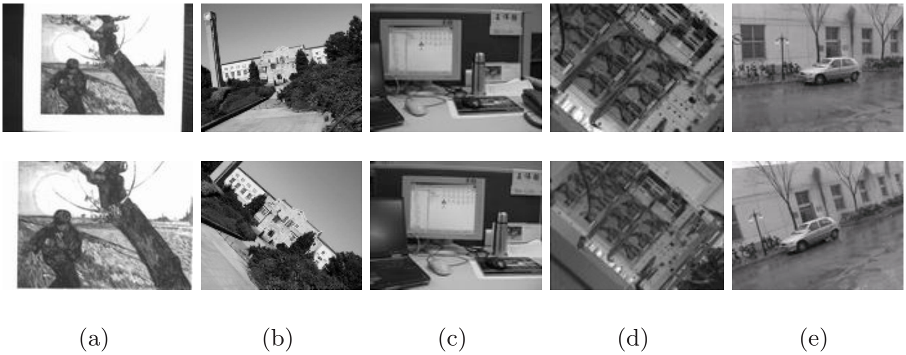


Fig. 6. The first row shows some of the query images. The second row shows the most similar images in the database, all of them are correct

5.2 Image Retrieval Application

We have evaluated the performance of SC descriptor in an image retrieval application. Experiments have been conducted for a small image database containing 40 images (including 8 real-world scenes). They show the robustness of our approach to image rotation, scale changes and small viewpoint variations.

The retrieval application is as follows. We first extract SC descriptors of each image in the image database. We compare each descriptor of a query image against all descriptors in the other image. Then we determine corresponding SC descriptors that are within a threshold. We regard the number of matched correspondences as a similarity between images.

Figure 6 shows the results of image retrieval experiments. The top row displays five query images. The second row shows the corresponding image in the database, which is the most similar one. The changes between the image pairs (first and second row) include scale changes, for example for pairs (a) and (d). They also include rotation changes, such as image pair (b). Furthermore, they include small viewpoint variations (image pair (c) and image pair (e)).

6 Conclusion

In this paper we propose a novel approach to solve the problem of obtaining reliable corresponding regions over two images in the presence of both large scale and rotation changes. We present a novel region descriptor based on the intensity distribution. Image matching and retrieval experiments show this descriptor is invariant to image rotation and scale changes as well as robust to limited changes in viewpoint.

Acknowledgements. This work is supported by National Hi-Tech Development Programs of China under grant No. 2003AA142140.

References

1. Schmid, C., Mohr, R.: Local Grayvalue Invariants for Image Retrieval. IEEE PAMI, 19 (1997) 530–534
2. Lowe, D.G.: Object Recognition from Local Scale-Invariant Features. In: ICCV, (1999) 1150–1157
3. Nagao, K.: Recognizing 3D Objects Using Photometric Invariant. In: ICCV, (1995) 480–487
4. Mikolajczyk, K.: Detection of Local Features Invariant to Affine Transformations. PhD thesis, INRIA, (2002)
5. Mikolajczyk, K., Schmid, C.: An Affine Invariant Interest Point Detector. In: ECCV, (2002) 128–142
6. Schaffalitzky, F., Zisserman, A.: Multi-view Matching for Unordered Image Sets, or “How Do I Organize My Holiday Snaps?”. In: ECCV, (2002) 414–431
7. Tuytelaars, T., Van Gool, L.: Wide Baseline Stereo Matching Based on Local Affinely Invariant Regions. In: BMVC, (2000) 412–425
8. Mikolajczyk, K., Schmid, C.: Indexing Based on Scale Invariant Interest Points. In: ICCV, (2001) 525–531
9. Baumberg, A.: Reliable Feature Matching across Widely Separated Views. In: CVPR, (2000) 774–781
10. Harris, C., Stephens, M.: A Combined Corner and Edge Detector. In: Proc. Alvey Vision Conf., Manchester (1988) 189–192
11. Xiao, J., Shah, M.: Two-Frame Wide Baseline Matching. In: ICCV, (2003) 603–609
12. Slater, D., Healey, G.: The Illumination-Invariant Recognition of 3D Objects Using Color Invariants. IEEE PAMI, 18 (1996) 206–210
13. Dufournaud, Y., Schmid, C., Horaud, R.: Matching Images with Different Resolutions. In: CVPR, (2000) 621–618
14. Lindeberg, T.: Feature detection with automatic scale selection. IJCV, 30 (1998) 79–116
15. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of Interesting Point Detectors. IJCV, 37 (2000) 151–172
16. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. May 2004. Preprint of article obtained from http://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf
17. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE PAMI, 24 (2002) 509–522

A Method for Blocking Effect Reduction Based on Optimal Filtering

Daehee Kim[1] and Yo-Sung Ho[2]

[1] Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong Yuseong-gu, Daejeon, 305-350, Korea
daeheekim@etri.re.kr
[2] Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong Buk-gu, Gwangju, 500-712, Korea
hoyo@gist.ac.kr

Abstract. In block-based coding schemes, the input image is segmented into small blocks that are processed independently; therefore, blocking effects occur along block boundaries. Various methods have been developed to reduce such blocking effects. In this paper, we propose a method for blocking effect reduction based on optimal filtering, and we compare its performance with those of others.

1 Introduction

The objective of image coding is to represent the image with as few bits as possible while retaining sufficient picture quality. Various image compression algorithms have been developed. Some of the more promising involve segmentation of the image into small subimages before coding. In this approach, the original image is divided into subimages, in most cases, square blocks of the equal size, and then each subimage is coded independently of the others. To reproduce the full image, the separated subimage blocks are reassembled by the decoder. The purpose of segmenting the image is to exploit local characteristics of the image and to simplify hardware implementation of the encoding algorithm. The transform coding is a typical example of the coding technique having image segmentation.

One of the fundamental problems of transform coding especially at the low bit rates is so-called the blocking effect. Since each block is processed independently, the reconstructed image at the decoder has discontinuities along block boundaries. This blocking effect is mainly due to independent quantization of transform coefficients in each block. Since the quantization takes place in the transform domain, the effect of quantization error is spread all over the spatial locations within the block. This phenomenon appears very annoying, as the coding bit rate decreases.

In order to reduce the blocking effect, various methods have been developed, such as the lapped orthogonal transform (LOT), the overlapping block method, the interleaving block method and post-filtering. However, each of those approaches has some drawbacks. The overlap method reduces blocking effects well

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 135–142, 2004.

without degrading image edges, but a major disadvantage of this method is an increase in the bit rate [1,2]. The post-filtering method is easy to implement and it works well. The post-filtering method does not increase the bit rate, but the filter degrades the edge content in the image [2]. The LOT is a popular method for reducing blocking effects. The optimal LOT has a major disadvantage of being highly sensitive to numerical errors, even with double-precision computations [3, 4]. The optimal LOT may not be easily factorable so that a fast algorithm may not exist [3]. The suboptimal LOT which has a fast algorithm, but the approximation for the suboptimal LOT is satisfactory only for the small block sizes. In our simulation for the LOT, we have shown the spread of the discontinuities along the block boundaries to adjacent blocks. The two-stage transform coding method [5] uses the total information of the image to reduce the blocking effect. In this algorithm, the error of each transform coefficient is spread to the entire image. Thus, the quality of image decreases exponentially as the bit rate decrease.

2 Optimal Filtering

In the previous section, we discussed blocking effect reduction algorithms that can be applied within local blocks or along local block boundaries. In this section, we develop a globally optimum filter instead of a locally optimum one. The globally optimum filter considers an entire image. Before we derive a globally optimum filter, let's consider a locally optimum filter to get a concept of the optimal filter.

First, we consider a block processing system with pre- and post-filters, as depicted in Fig. 1. One of the functions of the encoder is to shape the input signal spectrum into some appropriate form that takes into account quantization or noise degradations. At the decoder, an approximate inverse filter is employed to recover the original signal as much as possible.

In Fig. 1, the input noise **u** the quantization error **d** are stationary, uncorrelated, zero-mean random processes with known spectrum information. Here, the input and reconstructed signals **x** and $\tilde{\mathbf{x}}$ are vectors in the N-dimensional real space. We do not assume that **F** and **G** should be causal. We start by obtaining the optimal **G** for a given pre-filter; that allows us to derive an error expression that depends only on **F**. If we can find the pre-filter that minimizes the new error function, we can effectively obtain the jointly optimal filter pair. **D** and $\mathbf{D}^{-1}$ represent DCT and IDCT, respectively. We here assume that the input noise **u** is zero.

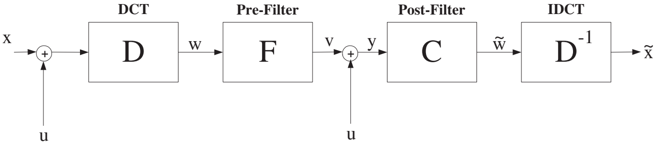


Fig. 1. Pre- and Post-Filter System

The pre-filter generates the intermediate signal $\mathbf{v}$ from $\mathbf{w}$, which is transformed by the matrix $\mathbf{F}$ and quantized through the quantizer. The post-filter builds an estimate $\tilde{\mathbf{w}}$ of $\mathbf{w}$, from which the final input estimate $\tilde{\mathbf{x}}$ is generated.

From Fig. 1, it is clear that

$$\mathbf{w} = \mathbf{D}\mathbf{x}, \quad \tilde{\mathbf{w}} = \mathbf{D}\tilde{\mathbf{x}} \tag{1}$$

Since we can assume the DCT and IDCT operations are lossless operations, the absolute mean-square error between $\mathbf{w}$ and $\tilde{\mathbf{w}}$ is the same as that between $\mathbf{x}$ and $\tilde{\mathbf{x}}$, that is given by

$$\xi_w = N^{-1}E[\|\tilde{\mathbf{w}} - \mathbf{w}\|^2] \tag{2}$$

Our problem, therefore, is reduced to find matrices $\mathbf{F}$ and $\mathbf{G}$ in Fig. 1 that minimize ξ_w. This is a typical classical problem in information theory, usually referred to as optimal block quantization or optimal block coding. We can make use of the cross-correlation between $\mathbf{v}$ and $\mathbf{d}$ to derive an expression for the error ξ_w as a function of the matrices $\mathbf{F}$ and $\mathbf{G}$. However, this would lead to matrix equations that are fairly difficult to manipulate. A much easier approach is to use the 'gain plus additive noise' model of scalar quantization. This model is derived by Malvar [3]. The quantizer output $\mathbf{y}$ is given by

$$\mathbf{y} = \boldsymbol{\Psi}\mathbf{x} + \tilde{\mathbf{d}} \tag{3}$$

where $\tilde{\mathbf{d}}$ is a noise source with no correlations, and $\boldsymbol{\Psi}$ is a diagonal matrix. The elements of $\boldsymbol{\Psi}$ depend on the autocorrelation $\mathbf{R}_{vv}$ [3]. With the relationship between v and y, we can modify the block diagram of Fig. 1 to Fig. 2.

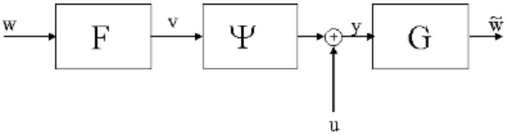


Fig. 2. Subsystem to be Optimized

We can rewrite (2) in the form

$$\begin{aligned}\xi_w &= N^{-1}tr\big\{E\big[(\mathbf{G}\boldsymbol{\Psi}\mathbf{F}\mathbf{w} + \mathbf{G}\tilde{\mathbf{d}} - \mathbf{w})(\mathbf{G}\boldsymbol{\Psi}\mathbf{F}\mathbf{w} + \mathbf{G}\tilde{\mathbf{d}} - \mathbf{w})^t\big]\big\} \\ &= N^{-1}tr\big\{\boldsymbol{\Lambda} + \mathbf{G}\boldsymbol{\Psi}\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^t\boldsymbol{\Psi}\mathbf{G}^t + \mathbf{G}\mathbf{R}_{\tilde{d}\tilde{d}}\mathbf{G}^t - 2\mathbf{G}\boldsymbol{\Psi}\mathbf{F}\boldsymbol{\Lambda}\big\}\end{aligned} \tag{4}$$

For any given $\mathbf{F}$, the optimal $\mathbf{G}$ can be obtained by setting $\partial\xi_w/\partial\mathbf{G} = 0$, which lead to

$$\mathbf{G}_{opt} = \boldsymbol{\Lambda}\mathbf{F}^t\boldsymbol{\Psi}\big(\boldsymbol{\Psi}\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^t\boldsymbol{\Psi} + \mathbf{R}_{\tilde{d}\tilde{d}}\big)^{-1} \tag{5}$$

For general cases of images coding, $\mathbf{F}$ is the identity matrix. If we assume that DCT and IDCT are lossless operations and DCT is suboptimal to KLT, we can expand the relation of (5) between the coefficients of the decoder and the encoder

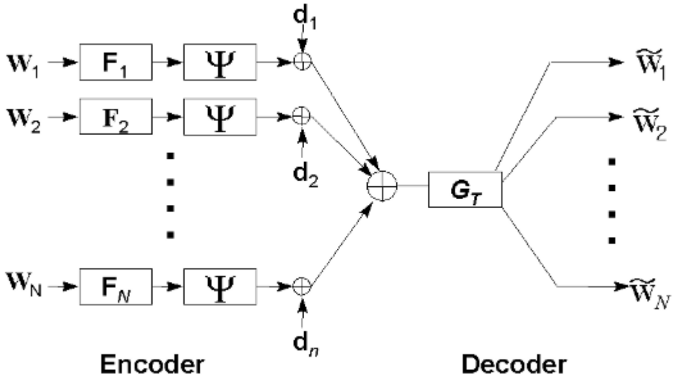


Fig. 3. Globally Optimum Filter $\mathbf{G}_T$

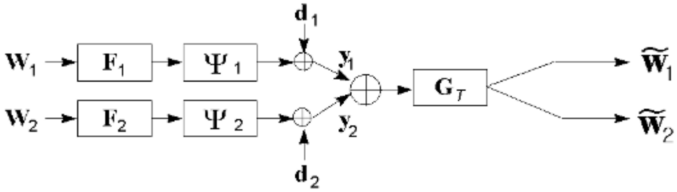


Fig. 4. Reduced System to be Optimized

to that between the original image and the reconstructed one. However, because of the independent block processing, we may have the blocking effect along the block boundaries in the reconstructed image.

In order to reduce the blocking effect, we should design a post-filter that is globally optimal for the entire image, instead of locally optimal for each block. For this work, we consider the system depicted in Fig. 3, where $\mathbf{w}_i$ is the input vector in the N-dimensional real space, $\mathbf{d}_i$ is the quantization error vector being uncorrelated with the $\mathbf{w}_i$, $\mathbf{F}_i$ is preprocessor, and $\boldsymbol{\Psi}_i$ is the diagonal matrix for quantization.

In this scheme, the entire signal is divided into small vectors, and each vector is independently processed in the encoder. At the decoder, we collect each coded vector to find a globally optimum filter $\mathbf{G}_T$.

Without loss of generality, we can simplify the derivation by considering only two blocks, as drawn in Fig. 4. The dimension of each block at the encoder is different from that of the globally optimum filter $\mathbf{G}_T$.

In order to manipulate each block and the global filter, we employ Kronecker product. We define the matrices, $\mathbf{K}_1$ and $\mathbf{K_2}$, for indicating each block, which is given by

$$\mathbf{K}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{K}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{6}$$

Also, we can write the entire original signal $\mathbf{x}$ and the entire reconstructed signal $\tilde{\mathbf{x}}$ in the form

$$\mathbf{x} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}, \quad \tilde{\mathbf{x}} = \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{bmatrix} \tag{7}$$

With (6), we can find the dimension extension of $\mathbf{w}_1$ and $\mathbf{w}_1$, as following:

$$\begin{bmatrix} \mathbf{w}_1 \\ 0 \end{bmatrix} = \mathbf{K}_1 \otimes \mathbf{w}_1, \quad \begin{bmatrix} 0 \\ \mathbf{w}_2 \end{bmatrix} = \mathbf{K}_2 \otimes \mathbf{w}_2 \tag{8}$$

Intermediate vectors, $\mathbf{y}_1$ and $\mathbf{y}_2$, and the quantization error vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, are represented by

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{w}_2 \end{bmatrix}, \quad \tilde{\mathbf{x}} = \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{bmatrix} \tag{9}$$

where

$$\mathbf{y}_i = \boldsymbol{\Psi}\mathbf{F}_i\mathbf{w}_i + \mathbf{d}_i \tag{10}$$

Similarly to the procedure for a locally optimum filter, we can define the error criterion to be minimized, which is given by

$$\xi_w = (2N)^{-1}\left\{E\left[\|\tilde{\mathbf{x}} - (\mathbf{K}_1 \otimes \mathbf{w}_1 + \mathbf{K}_2 \otimes \mathbf{w}_2)\|^2\right]\right\} \tag{11}$$

Using (8), (9), and (10), we can rewrite (11) as

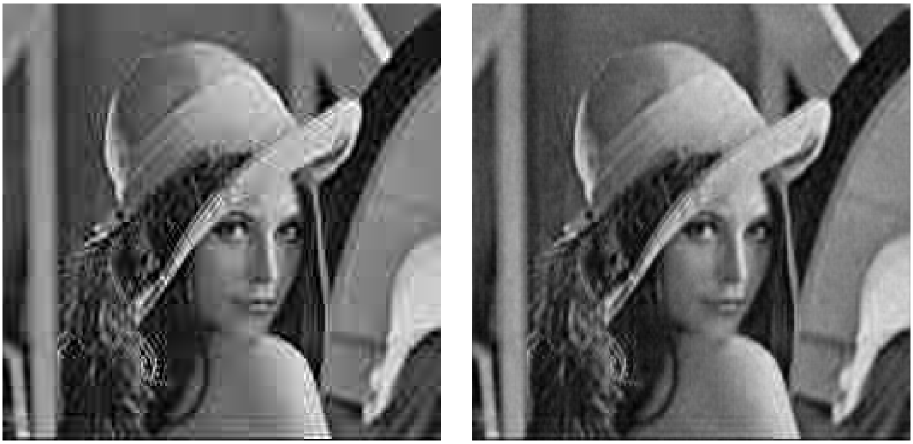
$$\begin{aligned}
\xi_w = (2N)^{-1}tr\big\{E\big[&\mathbf{K}_1 \otimes \mathbf{w}_1\mathbf{K}_1^t \otimes \mathbf{w}_1^t + \mathbf{K}_1 \otimes \mathbf{w}_1\mathbf{K}_2^t \otimes \mathbf{w}_2^t \\
&+ \mathbf{K}_2 \otimes \mathbf{w}_2\mathbf{K}_2^t \otimes \mathbf{w}_2^t + \mathbf{K}_2 \otimes \mathbf{w}_2\mathbf{K}_1^t \otimes \mathbf{w}_1^t\big] \\
&- 2\mathbf{G}_T E\big[\mathbf{K}_1 \otimes \mathbf{w}_1\mathbf{K}_1^t \otimes \mathbf{y}_1^t + \mathbf{K}_1 \otimes \mathbf{w}_1\mathbf{K}_2^t \otimes \mathbf{y}_2^t \\
&+ \mathbf{K}_2 \otimes \mathbf{w}_2\mathbf{K}_1^t \otimes \mathbf{y}_1^t + \mathbf{K}_2 \otimes \mathbf{w}_2\mathbf{K}_2^t \otimes \mathbf{y}_2^t\big] \\
&+ \mathbf{G}_T E\big[\mathbf{K}_1 \otimes \mathbf{y}_1\mathbf{K}_1^t \otimes \mathbf{y}_1^t + \mathbf{K}_1 \otimes \mathbf{y}_1\mathbf{K}_2^t \otimes \mathbf{y}_2^t \\
&+ \mathbf{K}_2 \otimes \mathbf{y}_2\mathbf{K}_1^t \otimes \mathbf{y}_1^t + \mathbf{K}_2 \otimes \mathbf{y}_2\mathbf{K}_2^t \otimes \mathbf{w}_2^t\big]\mathbf{G}_T^t \\
&= (2N)^{-1}tr\big\{\mathbf{G}_T\mathbf{R}_{yy}\mathbf{G}_T^t - 2\mathbf{G}_T\mathbf{R}_{xy} + \mathbf{R}_{xx}\big\}
\end{aligned} \tag{12}$$

where $\mathbf{R}$ represents a correlation function. For any given $\mathbf{F}_i$, assumed $\mathbf{F}_1 = \mathbf{F}_2$, the optimal $\mathbf{G}_T$ can be get by setting $\partial\xi_w/\partial\mathbf{G}_T = 0$, which leads to

$$\mathbf{G}_{Topy} = \mathbf{R}_{xy}(\mathbf{R}_{yy})^{-1} \tag{13}$$

Here, we have assumed that $\mathbf{F}$ is the identity matrix as in the local optimum filter, and we have assumed that DCT and IDCT are lossless operations. We can extend this relation to the original input signal at the encoder and the reconstructed signal at the decoder. By such an extension, we can see that (13) is in the form of the optimal Wiener filter. It is not strange because the Wiener filter is known as the optimal solution for many restoration problems.

Fig. 5(a) shows the reconstructed image before the blocking effect is reduced. Fig. 5(b) is the output image with the globally optimum filter. This technique



(a)Reconstructed Image (SNR 25.38 dB) (b)Processed Image (SNR = 26.44 dB)

Fig. 5. Results for LENA

has the best performance among all the methods discussed in this paper. This technique shows higher SNR about 1 dB than the other algorithms at 0.98 bpp.

We assumed the pre-filter is identity matrix. However, if we employ a pre-filter, we can get better performance than the scheme without the pre-filter. A disadvantage of this scheme is that we should know the information about the spectrum of the input signal.

3 Simulation Results

In this section, we compare various algorithms designed to reduce the blocking effect. For a fair comparison, each method should generate the same number of coding bits. It is important because some methods, such as overlap method, can generate more bits than other methods. Thus, the quantizer, with the bit allocation according to the variances of transform coefficients, can generate the same number of bits for various algorithms. Here, the optimized bit allocation table depends on the encoding algorithm, but the total number of bits in the bit allocation table should be independent of the employed algorithms. Since we do not use entropy coding, which is lossless coding, to make a fair comparison, our results have higher bit rate than those of standards.

Fig. 6 shows the SNR plot resulting from applying various algorithms to LENA. In Fig. 6, there are two reasons of the small difference in SNR. One is that SNR is not a good measure for the blocking effect, and the other is that we assign bits by the amount of energy. The globally optimum filter has the highest SNR value.

We define the discontinuity as the sum of absolute values of the differences taken along the block boundaries. The discontinuity along the block boundaries can represent the degree of the blocking effect. However, if we consider only the

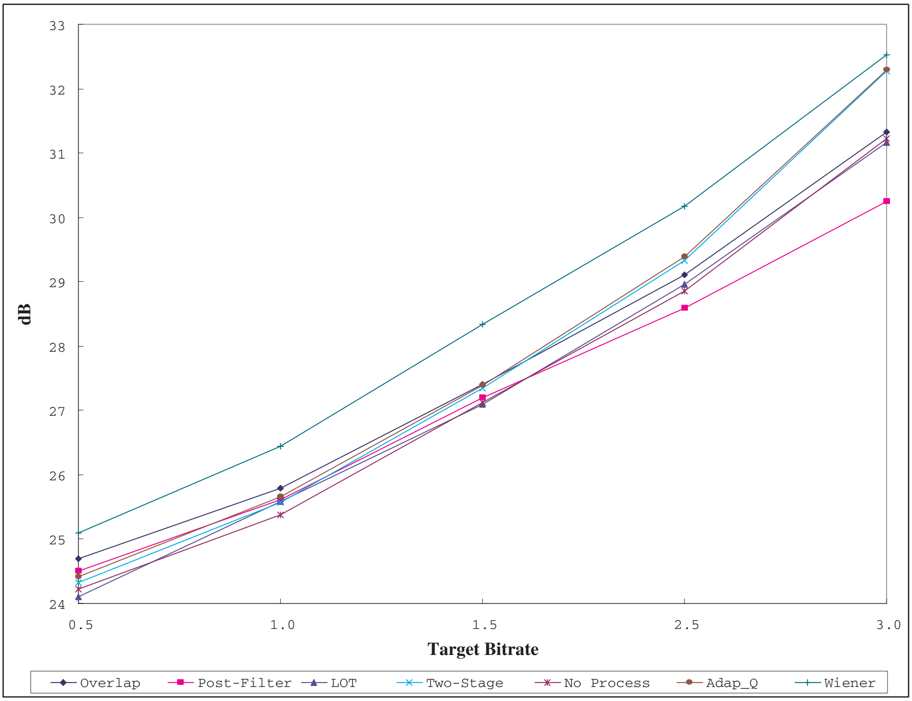


Fig. 6. Reduced System to be Optimized

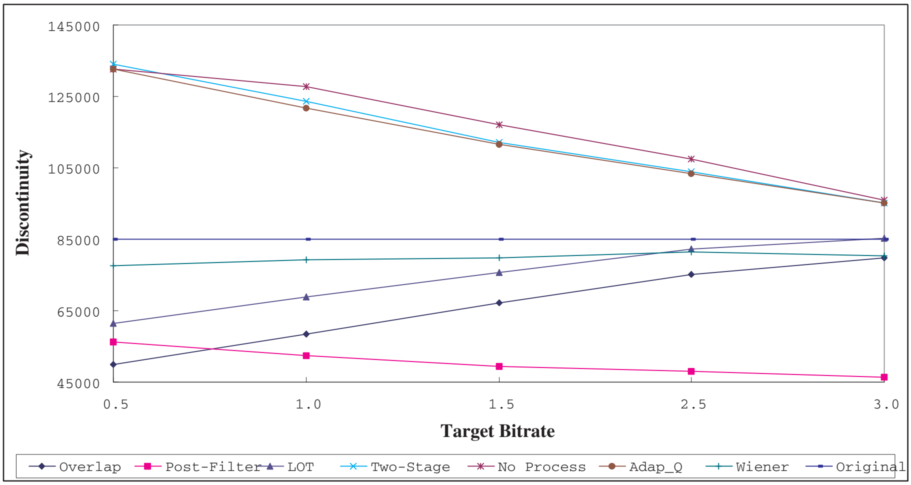


Fig. 7. Reduced System to be Optimized

method to minimize the discontinuity, some methods such as low-pass filtering can degrade the sharpness of the original image. While the HVS indicates the image from the overlap method is better than that from the modified overlap method, the discontinuity of the overlap method is larger than that of the modified overlap methods. Fig. 7 shows the discontinuities of various algorithms for LENA. The discontinuity of the reconstructed image depends on that of the original image. In this respect, we can say that the image has good quality as the discontinuity converges to that of the original image. In Fig. 7, the discontinuity of the original image is 85007 (flat line). In this respect, the globally optimum filter shows the best quality and the LOT shows the second.

4 Conclusions

In this paper, we have tested several algorithms for reducing the blocking effect. We have also have derived an optimal filter for reducing the blocking effect, assuming that we have an information of the input spectrum. The resulting post-filter is similar to the Wiener filter. If we use an estimation technique for the input spectrum, the performance of the Wiener filter may be degraded. In this paper, we have proposed a new criterion for comparing the degree of blocking effect reduction. In comparisons with this new criterion, our optimal filter shows the best result.

Acknowledgements. This work was 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. Reeve III, H.C., Lim, J.S.: Reduction of Blocking Effects in Image Coding. J. Optical Engineering, Vol. 23, No. 1 (1984) 34–37
2. Jarske, T., Haavisto, P., Defee, I.: Post-Filtering Methods for Reducing Blocking Effects from Coded Images. IEEE Int. Conf. Consumer Electronics (1994) 218–219
3. Malvar, H.S., Staelin, D.H.: The LOT: Transform Coding without Blocking Effects. IEEE Trans. Acoust., Speech, Signal Processing, Vol. 37, No. 4 (1989) 553–559
4. Cassereau, P.M., Staelin, D.H., Jager, G.D.: Encoding of Images based on a Lapped Orthogonal Transform. IEEE Trans. Communications, Vol. 37, No. 2 (1989) 189–193
5. Tran, A.: Block Effect Reduction in Transform Coding. SPIE Visual Communications and Image Processing, Vol.707 (1986) 182–187

Novel Video Error Concealment Using Shot Boundary Detection

You-Neng Xiao, Xiang-Yang Xue, Ruo-Nan Pu, Hong Lu, and Congjie Mi

Department of Computer Science and Engineering, Fudan University,
Shanghai 200433, P. R. China
{ynxiao,xyxue,rnpu,honglu,032021187}@fudan.edu.cn

Abstract. There are two conventional approaches for error concealment, i.e. the temporal approach and the spatial approach. Normally temporal error concealment is preferred due to its low computational complexity and good image quality after error concealment. However, the temporal error concealment will become ineffective for the frames at the shot boundary since the content of the frame is much different from the reference frame before the shot boundary. To address this problem, in the paper, we propose a novel error concealment method. First, a real time shot boundary detection method is proposed to detect whether the current damaged frame is at the shot boundary or not. Second, if the frame is at the shot boundary, the spatial error concealment method is adopted to conceal the damaged frame; otherwise, the temporal concealment method is adopted. Experimental results show that the proposed method is efficient, effective in fine concealed image quality, and accordingly quite suitable for kinds of video players.

1 Introduction

With the rapid development of digital TV, video conference, and computer network technology, application that concerns with video streaming in Internet is becoming more and more popular. However, since there is no special channel for video stream transmission in Internet, the transmission is very sensitive to channel disturbances. Besides, because of some compression techniques utilized in MPEG-1/2/4 standards, such as variable length coding (VLC) and temporal prediction, even a single bit error can lead to several frames being damaged and accordingly result in extremely poor visual quality.

Various methods have been proposed to solve the problems above at both sender and receiver. Some methods, such as data scalability, forward error correction (FEC) and rate control at the sender, aim at avoiding error occurrence. However, error concealment eases these problems by recovering the damaged regions caused by errors at the receiver [1, 2, 3, 4, 5, 6, 7].

Generally there are three types of error concealment methods. In forward error concealment methods [1], the decoder recovers the damaged regions relying on the redundancy added at the encoder. However, the redundancy may aggravate network congestion. In interactive error concealment methods [2, 3], the encoder adjusts its encoding policy according to feedback from the decoder. In such case, encoder and decoder need to be implemented by the same developer and thus it will limit the generality of video streams. Due to these disadvantages of the methods above, post processing based error

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 143–150, 2004.

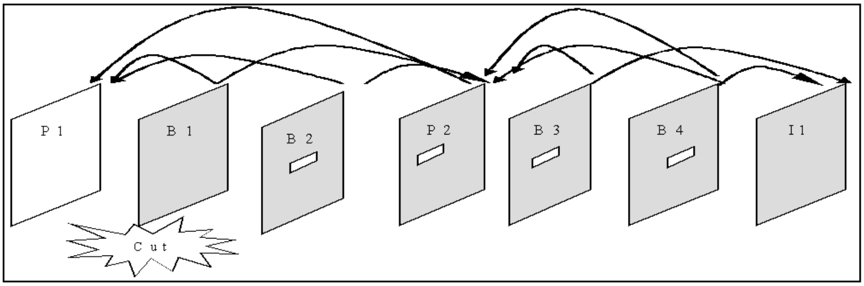


Fig. 1. Influence of shot change on temporal concealment

concealment is becoming popular [4, 5, 6, 7]. Taking into consideration the temporal correlation in video signals, temporal error concealment methods [4, 5] recover a damaged macroblock (MB) by its adjacent MBs in previous decoded frames. On the other hand, by making use of the spatial smoothness property of video signals, the spatial error concealment methods [6, 7] interpolate a damaged block from its adjacent blocks in the same frame.

Compared with spatial error concealment, temporal error concealment is normally preferred due to its low computational complexity and good concealed image quality. However, temporal error concealment will be ineffective when there are notable differences between the damaged frame and its reference frame, especially when these differences are caused by shot change. Figure 1 shows the influence of shot change on temporal error concealment, where a shot change occurs at B1. It is obvious that the visual quality will be severely bad if damaged MBs in its following frames are concealed by MBs in P1, which is in the previous shot.

To address the problem above, in this paper, we propose an error concealment method using shot boundary detection. Directly utilizing the information acquired during decoding process, such as motion vectors of P, B-frames and direct current coefficients (DC) of I-frame, a real time shot boundary detection method is proposed to judge whether or not the damaged frame is a shot boundary. In the case of shot boundary, the damaged frame is concealed by an improved Spatial Block-based Split-Match error concealment method [4]; otherwise a Temporal Forward-Backward Block-Matching error concealment method [4] is performed.

The remainder of this paper is organized as follows. In section 2, we describe our proposed error concealment method. Section 3 presents experimental results to show the effectiveness of our proposed method. And the paper is concluded in Section 4.

2 Description of Proposed Method

In this section, we first introduce the error detection method and spatial/temporal error concealment methods adopted in our method. Second, we propose the fast shot boundary detection method on compressed video directly. Finally, we propose the hybrid spatial/temporal error concealment method using real time shot boundary detection. Furthermore, though our proposed method is applicable to MPEG-1/4 and other video coding standards, for easy of presentation, we take MPEG-2 coded video stream as example.

2.1 Error Detection

Error detection is the fundamental stage for error concealment, which aims at finding out whether an error has occurred or not, and if it has occurred, where it is.

As a MB is encoded by VLC in MPEG-2, an error in a MB is prone to influence the following MBs in the same slice till the next slice. Moreover, since the first and last MBs of a slice are in the same horizontal row, the damaged MB can only influence MBs in the same horizontal row. Hence we use the pixel differences between the boundary pixels in current MB and its two vertically neighboring MBs to detect errors. Moreover, as the difference of U component is most notable among that of Y, U, V components based on extensive experiments, we regard the difference of U component's gradient as the discrepancy criterion of a MB and its two vertically neighboring MBs.

Let TOP be sum of gradients between the last two lines of the top neighboring MB, i.e.

$$TOP = \sum_{k=0}^{15} |MB_{i-1,j}(15,k) - MB_{i-1,j}(14,k)|,$$

and TOP_CUR be sum of gradients between the first line of the current MB and the last line of the top neighboring MB, i.e.

$$TOP\_CUR = \sum_{k=0}^{15} |MB_{i,j}(0,k) - MB_{i-1,j}(15,k)|,$$

where $MB_{i,j}(m,n)$ represents the U component's value at pixel (m,n) of $MB(i,j)$. Then BOT and BOT_CUR can be defined in the similar way.
Normally TOP, BOT, TOP_CUR and BOT_CUR are small due to the smoothness of video frame. However, in case of error occurrence, TOP_CUR, BOT_CUR will become large. Figure 2 shows an example.

To avoid false alarms, only when the following two formulas

$$TOP\_CUR > TOP \times TH\_U; \; BOT\_CUR > BOT \times TH\_U$$

are satisfied simultaneously, the current MB is declared to be damaged. In our implementation, the threshold TH_U is set to 3.

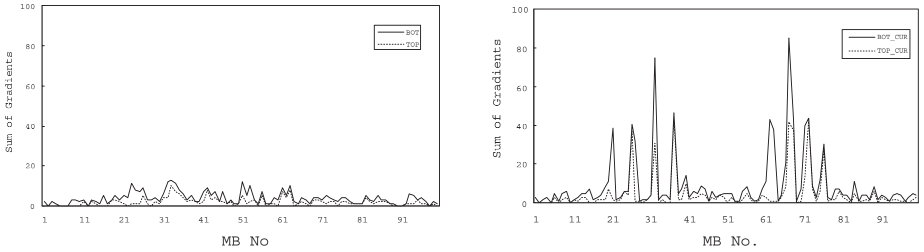


Fig. 2. Comparison of TOP, BOT and TOP_CUR, BOT_CUR in case of errors occurrence

2.2 Spatial/Temporal Error Concealment

For spatial error concealment, we adopt improved Spatial Block-based Split-Match error concealment method [4]. Firstly, a large region (denoted as A) in damaged MB is chosen to search the matched region in its vertically neighboring MBs. If there is no such matched region, A is spitted into two smaller ones and the match searching is performed for the spitted blocks. The steps are performed iteratively until the entire MB is matched. Then concealment is performed by copying the matched region to the damaged region.

For temporal error concealment, we adopt Temporal Forward-Backward Block-Matching error concealment method [4]. Specifically, suppose current damaged MB is MB_C, and its neighboring top and bottom MBs are MB_A and MB_B, respectively. In MB_C's reference frame we search the best-matched regions MB'_A and MB'_B for MB_A and MB_B, then use the region located between MB'_A and MB'_B to conceal MB_C.

2.3 Fast Shot Segmentation in Compression Domain

There are lots of methods for shot boundary detection [8] and the shot boundaries can be classified as dissolve, fade in/fade out and shot cut. Generally, dissolve and fade in/out can last several frames, so the concealment effect will be acceptable even if we conceal an error MB at the shot boundary by temporal error concealment method. Hence, in our paper, we focus our attention on shot cut detection. Since here the key of cut detection is speed not precision, we only utilize information directly obtained from decoding process. So the shot boundary detection method will only cost a few computation and memory resources.

Based on the principle of coding, MPEG-2 encoder will choose the most proper coding mode (intra, forward or backward prediction) during encoding. For instance, when a shot cut occurs, inter-prediction will fail and intra-coding is the best choice.

First, for P-frame, let R_P denote ratio of the number of MBs without motion vector ($|MB_{non}|$) and the number of those with forward motion vector ($|MB_F|$), i.e.

$$R_P = \frac{|MB_{non}|}{|MB_F|}.$$

In general, as most MBs in P-frame will choose forward prediction mode to reduce bitrate, R_P is small. However, when a shot change occurs at a P-frame, because of large differences between the current frame and its forward reference frame, most MBs in P-frame will choose intra-coding mode, so R_P is large. Thus we can detect shot boundary at P-frame by examining R_P.

For B-frame, let R_B denote ratio of the number of MBs with backward motion vector ($|MB_B|$) and the number of those with forward motion vector ($|MB_F|$), i.e.

$$R_B = \frac{|MB_B|}{|MB_F|}.$$

Generally, as most MBs in B-frame will use bi-prediction coding mode to reduce bitrate, R_B fluctuates around 1. When a shot change occurs at a B-frame, most MBs in B-frame use backward prediction, so R_B is large. Thus we can detect the shot boundary at B-frame by examining R_B.

In this paper, we use sliding windows to detect shot boundary at P-frame and B-frame. For R_P and R_B, we set their own slide windows respectively. Here we take the sliding window of R_P for example.

Let W be a sliding window with length l, $w_i (i \in [0, l))$ an element of W (namely value of R_P), $E(W)$ the expectation of W, $\sigma(W)$ the mean square deviation of W, $FrameNo$ the frame No of a P-frame, $ShotCount$ the shot counter.

Sliding Window Algorithm:

```
1. Set $FrameNo$=0, $ShotCount$=0.
2. For each element of the window, set $w_i$=InitialEle, $(i \in [0, l))$, $i = 0$, and compute $E(W)$ and $\sigma(W)$.
3. If $|R_P(FrameNo) - E(W)| > TH_P \times \sigma(W)$, goto step 4; otherwise goto step 5.
4. $Shot[ShotCount] = FrameNo$, $ShotCount = ShotCount + 1$, $FrameNo = FrameNo + 1$, goto step 3.
5. $W(i \ mod \ l) = R_P(FrameNo)$, recompute $E(W)$ and $\sigma(W)$, $i = i + 1$, $FrameNo = FrameNo + 1$. If last P-frame is reached, quit; otherwise goto step 3.
```

In this paper, for P-frame, we set InitialEle=1000, the length of sliding window $l \in [8, 10]$, and TH_P=4; for B-frame, InitialEle=1000, $l \in [18, 22]$, and TH_B=6.

On the other hand, the possible shot change at I-frame can be detected by the following method.

There are 64 DCT coefficients in each block (8×8 pixels) of a MB, the first of which is DC coefficient (denoted as $c(0, 0)$). Then $c(0, 0) = \frac{\sum_{x=0}^{7} \sum_{y=0}^{7} c(x,y)}{64}$, namely, DC coefficient is the mean of the whole block. All the DC coefficients in a frame constitute the DC frame of the original frame. Though the resolution of the DC frame is 1/64 of that of the original frame, it maintains most frame texture information.

We use DC frame of luminance in I-frame to detect cut change. Let I_i and I_{i+1} be two neighboring I-frames, then the difference between their DC frames is defined as:

$$Diff\_I\_DC(I_i, I_{i+1}) = \sum_{k=0}^{TotalBlock-1} |DC_k^i - DC_k^{i+1}|$$

There will be a great peak of $Diff\_I\_DC(I_i, I_{i+1})$ in case of shot cut. Similarly sliding window can also be used for $Diff\_I\_DC(I_i, I_{i+1})$, where, $l = 4$, the threshold $TH\_I = 3$, and $InitialEle = \frac{width \times height \times 8}{256}$.

Since the thresholds in our detection method are set low, the shot boundary detection method proposed can achieve high recall rate, while sometimes it will lead to false alarms. As these alarms only make the decoder choose spatial concealment instead of temporal concealment, the influence can be neglected.

2.4 Error Concealment Method Using Shot Boundary Detection

Our error concealment method using shot boundary detection is as follows:

Step 1: Set $n = 1$.
Step 2: Decode the nth frame.
Step 3: Detect whether there are errors in the nth frame. If an error occurs, goto Step 4; otherwise goto Step 6.

Step 4: Detect whether the nth frame is a shot boundary.
Step 5: If the nth frame is a shot boundary, spatial error concealment is performed; otherwise temporal error concealment is performed.
Step 6: If the nth frame is the last one, quit; otherwise $n = n + 1$, goto Step 2.

3 Experiment Results

3.1 Experiment Setup

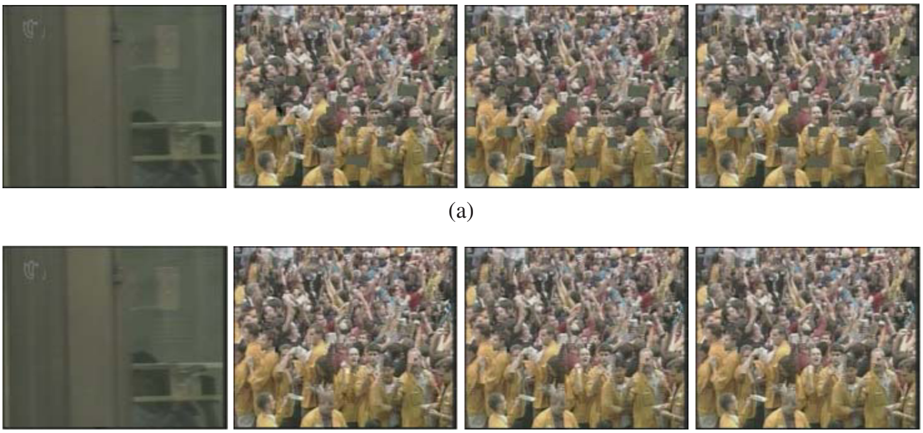
The data used in our experiment are digital TV programs captured by DVB card from satellite, whose data format is MPEG-2 TS stream. The bit rate is about 4.5∼6 Mbps, and video format is main profile/main level. When transmitting over networks, the data uses RTP over UDP, with the RTP packet size 1128 byte. The video programs are from TVBS, BBC, and Music Asia stations, respectively. In our experiments we compare our proposed error concealment method using shot boundary detection (ECSBD) with the Simple Block-Replacing Temporal error concealment method without shot boundary detection (SBRT) [9] and the Temporal Forward-Backward Block-Matching error concealment method without shot boundary detection (TFBBM) [4].

3.2 Experiment Results

To evaluate our method, we only compute the mean of PSNRs of several frames at the damaged shot boundary, while frames inside a shot are not taken into account. The experimental results are tabulated in Table 1. It can be observed from Table 1 that our method adaptively switches between temporal and spatial error concealment at shot boundary, good image quality can be achieved. When the network packet loss rate is 1×10^{-3}, 5×10^{-3} and 1×10^{-2}, PSNR of our method is higher than that of SBRT by 1.02, 1.01, and 1.18, and higher than that of TFBBM by 0.64, 0.60, and 0.72, respectively. While when an infrequent false alarm of shot boundary occurs, PSNR of our method will be lower than that of TFBBM by 0.2-0.4 because of the adoption of spatial error concealment.

Table 1. Comparison in PSNR of three methods with different packet loss rate

Sequence	Number	Packet Loss	SBRT	TFBBM	ECSBD
	of Shot	Rate			
TVBS	27	1×10^{-3}	29.93	30.26	30.89
		5×10^{-3}	28.17	28.62	29.34
		1×10^{-3}	24.35	24.93	25.76
BBC	21	1×10^{-3}	32.34	32.73	33.29
		5×10^{-3}	31.65	31.97	32.53
		1×10^{-3}	29.17	29.61	30.42
Music Asia	32	1×10^{-3}	26.71	27.12	27.86
		5×10^{-3}	25.57	26.03	26.54
		1×10^{-3}	21.87	22.21	22.75



(a)

(b)

Fig. 3. Error concealment effect of TFBBM (a) and ECSBD (b)

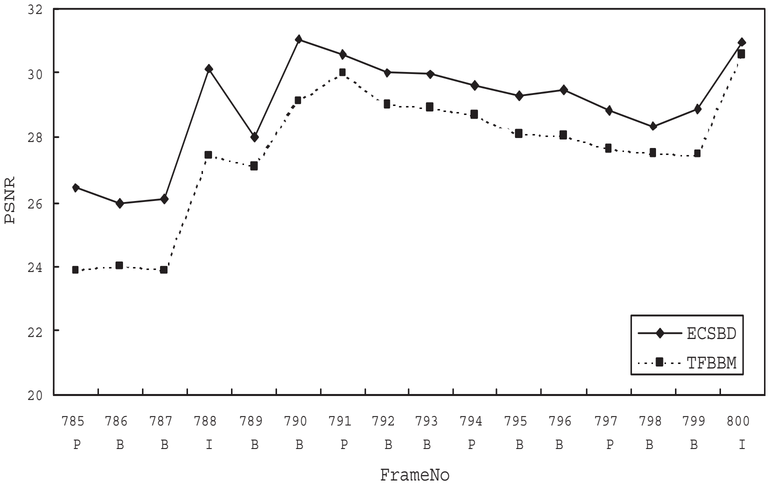


Fig. 4. Comparison in PSNR at a shot boundary: ECSBD vs. TFBBM

Figure 3 shows the subjective comparison of our proposed method, i.e. ECSBD, and TFBBM at a shot boundary. It can be visually observed that our proposed method achieves better performance than TFBBM. Specifically, the mosaic blocks in Fig. 3A are concealed by our proposed method.

Besides, we find that if cut shot occurs at a frame, its following P-frame and I-frame can also detect this cut and use spatial error concealment method to avoid error propagation. Figure 4 shows PSNR of a damaged frame (the P-frame with frame No 785) at a shot boundary and its following several frames, varying with frame No. In Fig. 4, both the P-frame with frame No 785 and the I-frame with frame No 788 detect the shot boundary by their own sliding window algorithms and then adopt spatial error concealment, respectively. So the error will not propagate after the I-frame and the following frames can gain high PSNR. However, as TFBBM always uses temporal error concealment, the damaged frame with frame No 785 is concealed using its reference frame in previous shot. So the error propagates to its following frames until another I-frame in the next GOP. For this instance, the average PSNR for all the frames showed in Figure 4 of our method is higher than that of TFBBM by 1.39.

4 Conclusion

In this paper, we propose an error concealment method using shot boundary detection. Making use of the information acquired during decoding process, we propose a real time shot boundary detection method to judge whether or not the damaged frame is at shot boundary, then adaptively make a choice between spatial error concealment and temporal error concealment. Experimental results show that our proposed method can achieve better performance in comparison with globally adopting a spatial or temporal error concealment method.

Acknowledgements. This work was supported in part by Natural Science Foundation of China under contracts 60003017 and 60373020, China 863 Plans under contract 2002AA103065, and Shanghai Municipal R&D Foundation under contracts 03DZ15019 and 03DZ14015, MoE R&D Foundation under contract 104075.

References

1. Y. Wang and S. Lin.: Error-Resilient Video Coding Using Multiple Description Motion Compensation. IEEE Trans. on Circuit and System for Video Technology. **12(6)** (2002) 438–452
2. T. Wiegand, N. Färber, K. Stuhlmüller, and B. Girod.: Error-Resilient Video Transmission Using Long-Term Memory Motion-Compensated Prediction. IEEE Journal. on Selected Areas in Communications. **18(6)** (2002) 1050–1062
3. C.-S. Kim, R.-C. Kim and S.-U. Lee.: An Error Detection and Recovery Algorithm for Compressed Video Signal Using Source Level Redundancy. IEEE Trans. on Image Processing. **9(2)** (2000) 209–219
4. S. Tsekeridou and I. Pitas.: MPEG-2 Error Concealment Based on Block-Matching Princ-iples. IEEE Trans. on Circuit and System for Video Technology. **10(4)** (2000) 646–658
5. M.-J. Chen and S.-Y. Lo.: Temporal Error Concealment Using Two-Step Block Matching Principle. Intl. Conf. on Consumer Electronics. (2001) 172–173
6. L.-D. Soares and F. Pereira.: Spatial Shape Error Concealment for Object-Based Image and Video Coding. IEEE Trans. on Image Processing. **13(4)** (2004) 586–599
7. J.-W. Suh and Y.-S. Ho.: Error Concealment Techniques for Digital TV. IEEE Trans. on Broadcasting. **48(4)** (2002) 299–306
8. X.-B. Gao and X.-O. Tang.: Unsupervised Video-Shot Segmentation and Model-Free Anchorperson Detection for News Video Story Parsing. IEEE Trans. on Circuit and System for Video Technology. **12(9)** (2002) 765–776
9. J. F. Arnold, M. R. Frater and J. Zhang.: Error Resilience in the MPEG-2 Video Coding Standard for Cell Based Networks-A Review. Signal Process: Image Communication. **14** (1999) 607–633

Using Only Long Windows in MPEG-2/4 AAC Encoding

Fu-Mau Chang and Shingchern D. You

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-2/4 AAC standard uses both long and short windows. However, using short windows complicates the implementation of encoders. In this paper, we propose a simple method to encode signals only with long windows. By setting all the scalefactors to be the same, the proposed approach is fully compatible to the standard. The objective experiments using PEAQ show that the proposed approach has a better coded quality at 64 kbps for mono music. Subjective listening experiments show that the proposed approach performs much better for coding mono music at a lower bitrate of 32 kbps. Therefore, the proposed approach is a promising alternative in coding transient signals without using short windows.

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 development of MPEG-4 natural audio coding [6] was largely based on the AAC coding scheme. In addition, AAC is also getting more popular for commercial use. For example, musical tracks (songs) sold in the Apple's web[12] are coded using the AAC format.

In the AAC standard, a certain number of PCM samples in a channel, depending on the signal type, are multiplied by a window function. Then, the Modified Discrete Cosine Transform (MDCT) is applied to the windowed results for time-to-frequency conversion (or subband analysis) [13]. To achieve perfect reconstruction (PR), the second half of the PCM samples covered in the previous window are in the first half of the current window scope. Four window types are

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 151–158, 2004.

used in the AAC, namely long window, short window, start window, and stop window. Long windows are used for stationary signals to achieve higher coding gain. On the other hand, short windows are used for transient signal for better time resolution. In order to smoothly change the window type from a long window to a short window, an intermediate window called start window is used. Also, a stop window is used for switching from a short window to a long window. A long window covers 2048 samples which are converted to 1024 spectral lines after the MDCT operation. These spectral lines, after quantization, are packed as a block in the bitstream. In the case of short windows, 2048 PCM samples are covered by eight consecutive short windows to obtain eight sets of 128 spectral lines to be packed in a coded block. Both start and stop windows cover 2048 PCM samples, therefore the windowed samples are transformed to 1024 spectral lines. The psychoacoustic model in the encoder determines whether the signal in the present block is stationary or transient based on a measure called Perceptual Entropy (PE). If the PE is greater than a threshold in a block, then this block is to be coded using short windows. Otherwise, the block is encoded using a long window.

2 The Issue of Using Short Windows

The MPEG audio standards use different types of window to provide a mechanism to switch between a higher coding efficiency and a higher time resolution. However, this type of design makes the encoder (as well as the decoder) more complicated. For example, in order to use short windows, the psychoacoustic model has to calculate the Signal-to-Masking Ratio (SMR) for both long and short windows. Based on our experiments, if the codes related to short windows are deleted, then the psychoacoustic model can run three times faster. Moreover, the program of the window switching part contributes about 15 % of the total code size in the ISO's reference software. Therefore, in terms of implementation, it is advantageous to use only long windows in the coding process.

The first reason for using short windows is to reduce the pre-echo noise. However, since the Temporal Noise Shaping (TNS) tool in the AAC can also be used to control pre-echoes [7], this reason is not justified. The second reason is to provide a higher time resolution for transient signals. Higher time resolution is achieved by using one set of scalefactors per window so that scalefactors (and the corresponding quantization step sizes) can be changed in a shorter time instance. As a transient signal usually has a rapid change in waveform, coding such a signal may require to use different scalefactors in a shorter period of time. Unfortunately, using short windows results in lower coding gain and higher overhead. This in turn degrades the quality of coded signal at lower bitrates. On the other hand, if the coding bitrate is higher enough, it is not important whether using short windows to encode transient signals or not because the quantization noise can be controlled to an ignorable level. Therefore, using short windows in the encoding process should be careful.

3 The Proposed Method

Previously we have proposed an approach to use only long windows in AAC encoding [8]. Closely examine the functionality of that approach, we found that the number of bits used in coding each scalefactor band was roughly the same for transient signals. This gives us a hint. For transient signals, we may assign the number of bits to a scalefactor band solely based on the relative energy contained in the corresponding scalefactor band. In other words, the scalefactors in every band are assigned to be equal to the global gain, and the rate control is achieved by varying this value. Again, recall that this type of arrangement is carried out if the psychoacoustic model indicates that a short windows is to be used. The complete encoding process of the proposed approach is then as follows. The input PCM samples are passed through the psychoacoustic model to calculate the perceptual entropy (PE). If the PE value is less than a threshold, then the current block is encoded with a normal (conventional) encoding process. On the other hand, if the PE value is greater than the threshold, indicating the transient nature of the signal, then a rate control routine is executed with all scalefactors set to be the same as the global gain.

4 Experiments and Results

We implement the proposed approach based on the ISO's reference software because ISO's source code is easy to obtain, read, and modify. The quantization part of the program is modified in such a way that all scalefactors are set to the same as the global gain if PE is greater than a threshold (+400). Other parts of the software remains unchanged. In the following, the comparison counterparts are also (slightly) modified from the ISO's reference software. Since all comparison methods are based on the ISO's software, in terms of comparison they are on the same ground.

In our previous paper [8], we show that a transient signal, as given in Fig. 1(a), is distorted if it is coded using long windows only by setting the PE threshold of the ISO's reference program to infinity, as given in Fig. 1(b). Therefore, the first experiment was to encode the same transient signal using the proposed approach to see if coded signal is still distorted. The result, as given in Fig. 1(c), show that the obvious distortion is no longer appear. This give us some confidence that this approach is promising in removing an obvious distortion.

The quality assessment experiments were conducted in two parts. The first part used the EAQUAL [11], a PEAQ [9,10] program, as the objective quality measurement, and the second part was listening (subjective) experiments. Since the PEAQ is mainly used to score coded music with high quality, musical signals coded lower bitrates are scored by subjective (listening) tests.

4.1 Results of PEAQ Tests

The first part of experiments used about one hundred pieces of various types of music. The types of music include classical music, soft music, hard rock and

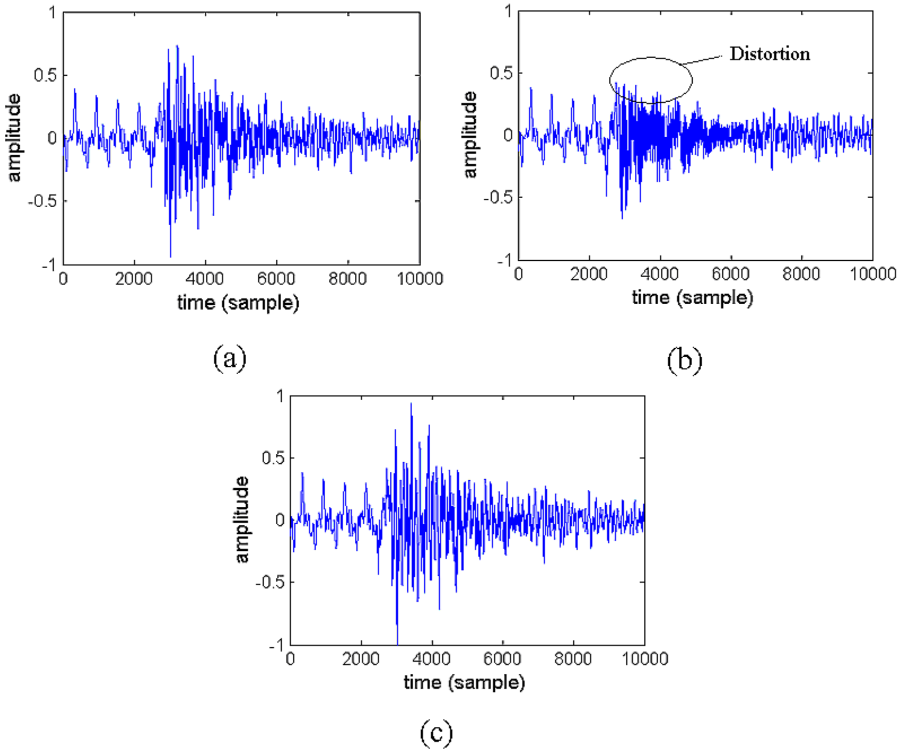


Fig. 1. (a) The transient signal under test; (b) The coded results using only long windows; (c) The coded results using the proposed approach. The distortion is no longer apparent.

electronic music. The musical pieces are originally recorded in mono with a sampling rate of 44.1 ks/s. These pieces are then encoded with a bitrate of 64 kbps. The comparison counterparts to the proposed approach are the following: (i) the ISO's reference program without modification, i.e., the PE threshold set to -1000 (called as ISO_org); (ii) the ISO's approach with the PE threshold set to +400 (called ISO_400); (iii) a long-window-only approach by setting to infinity the PE threshold of the ISO's program (called ISO_long). The method given in [8] was excluded from the comparison list for its incompatibility to the ISO's standard.

The average ODG of the mentioned methods are given in Table 1. In addition, the relative performance of the proposed approach versus one of the comparison

Table 1. Average ODG of various approaches.

Approach	ISO_org	ISO_400	ISO_long	Proposed
AVG ODG	-2.02	-1.70	-1.74	-1.47

counterparts are plotted in Figs. 2 to 4. In the figures, the horizontal axis is the ODG differences between the proposed approach and one counterpart. The positive side of the horizontal axis represents that the proposed approach has a higher ODG. The vertical axis represents the number of songs in each specified range of ODG difference. Based on the figures it can be seen that the proposed approach has a better ODG over other methods on most of the music tracks under test. In addition, the results are consistent to our previous claim [8]: The default value of PE threshold (-1000) of the ISO's reference program is not a good choice. A better value for the threshold would be +400.

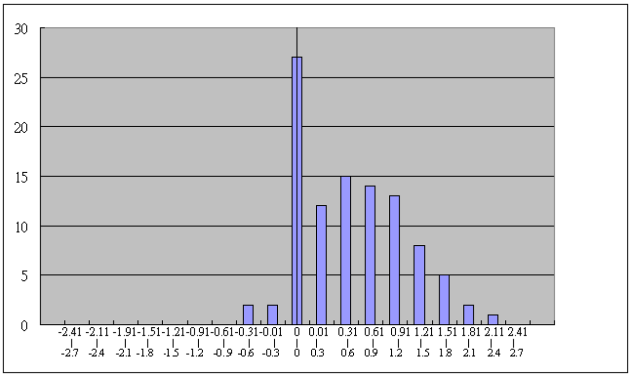


Fig. 2. The ODG difference between the proposed approach and ISO_1000. The positive number on the horizontal axis means that the proposed approach is better. The vertical axis is the number of songs.

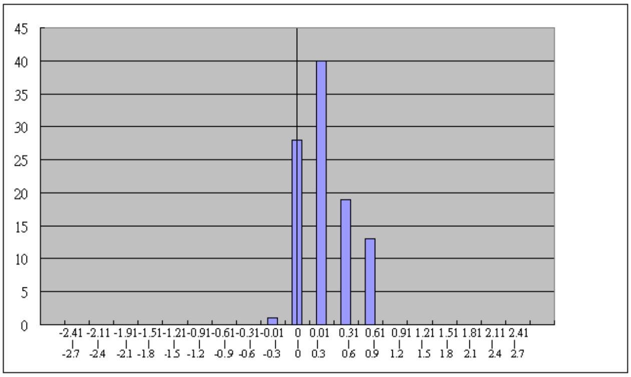


Fig. 3. The ODG difference between the proposed approach and ISO_400. The positive number on the horizontal axis means that the proposed approach is better. The vertical axis is the number of songs.

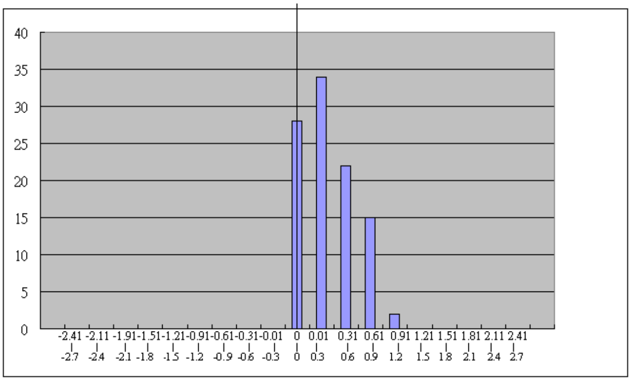


Fig. 4. The ODG difference between the proposed approach and ISO_long. The positive number on the horizontal axis means that the proposed approach is better. The vertical axis is the number of songs.

4.2 Results of Subjective Experiments

As we mentioned previously, the PEAQ is not intended to evaluate the sound quality at low bitrates. To fully explore the limitation of and to subjective evaluate the coded quality of the proposed approach, we also carried out the listening experiments. The coding bitrates were set to 32 kbps and 64 kbps. The comparison counterpart was the ISO_400 approach. Due to lacking of experienced audiences, we used a simplified CMOS (Comparative Mean Opinion Score) method in the experiments. 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 opinion is: A is better than B, A is equal to B, or A is worse than B. The signal coded by the proposed method was randomly assigned to either A or B. Besides, the audiences had no knowledge about which one was coded by the proposed method. The signals for comparison were eight pieces of music containing many strong attacks (transient signals). The contents of the signals are listed in Table 2. The experimental results are given in Table 3 and 4. Since our audiences are not trained experts, they are unable to judge the quality difference at 64 kbps. Based on the subjective results, it is clear that the proposed approach also has a better quality at lower bitrates. Overall speaking, the proposed approach is better for most of the test soundtracks from low to high bitrates. Moreover, because the proposed approach does not use short windows, it has the implementation advantage.

Table 2. Signals used in the experiments.

else	Soft music (female vocal)	michael	Rock and Roll
heavy	Heavy metal music	sampg	Soft music (male vocal)
castanets	castanets	harp	Harpsichord
eagle	Pop music	song1	pop music

Table 3. The CMOS scores between the proposed approach and the ISO_400 method at 32 kbps.

Music name	Proposed approach better	Both equal	ISO_400 better	Average score
else	6	6	3	0.20
heavy	10	4	1	0.60
castanets	11	3	1	0.67
eagle	9	4	2	0.47
michael	8	4	3	0.33
sampg	7	4	4	0.20
harp	2	11	2	0.00
song1	1	14	0	0.07

Table 4. The CMOS scores between the proposed approach and the ISO_400 method at 64 kbps.

Music name	Proposed approach better	Both equal	ISO_400 better	Average score
else	4	8	3	0.07
heavy	5	6	4	0.07
castanets	4	7	4	0.00
eagle	5	7	3	0.13
michael	1	13	1	0.00
sampg	1	12	2	-0.07
harp	3	10	2	0.07
song1	1	13	1	0.00

5 Conclusions

In this paper, we have demonstrated that using only long windows in AAC encoding does not affect the sound quality of the coded music in most instances. On the contrary, it improves the quality in many cases if using the proposed approach. Since the proposed approach uses only rate-control loop in coding transient signals, it is fully compatible with the AAC standard. Compared with the conventional AAC encoding scheme, the proposed approach reduces the complexity of the encoder, both in computation and structure aspects.

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. Herre J., Johnston J. D.: Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS). 101st Conference of the Audio Engineering Society, Los Angeles, CA (1996) pre-print 4384
8. Yu C-H, You S. D.: On the Possibility of Only Using Long Windows in MPEG-2 AAC Coding. Lecture Notes on Computer Science, LNCS 2532, Springer-Verlag, (2002) 663–670
9. ITU: Methods for Objective Measurements of Perceived Audio Quality. ITU-R Rec. BS. 1387 (1998)
10. Thiede T., et al: PEAQ-The ITU Standard for Objective Measurement of Perceived Audio Quality. Journal of Audio Eng. Soc. 48 (2000) 3–29
11. Available at www.mp3-tec.org/programmer/misc.html.
12. Please refer to www.apple.com/itunes for details.
13. Princen J. P., Johnson A. W., Bradley A. B.: Subband Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation. Proc. IEEE ICASSP, Dallas, TX, USA, (1987) 2161–2164

Frequency Weighting and Selective Enhancement for MPEG-4 Scalable Video Coding

Seung-Hwan Kim and Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea
{kshkim,hoyo}@gist.ac.kr

Abstract. In MPEG-4 scalable video coding, only a small portion of input data is coded in the base layer, and most signal components remain in enhancement layers. In this paper, we propose a new frequency weighting method to send more sensitive frequency coefficients faithfully with respect to the human visual system (HVS). In order to implement the frequency weighting method by bit-plane coding, we obtain a frequency shift matrix from the HVS-based frequency weighting matrix. We also propose a fast selective enhancement method using coding information, such as motion vectors and residual image blocks. By applying the proposed ideas, we have improved visual quality of reconstructed images. In order to measure subjective image quality appropriately, we define a new error metric, called as the just noticeable difference error (JNDE), based on the Weber's law.

Keywords: FGS, Frequency weighting, Selective enhancement, JNDE

1 Introduction

Recently, several scalable video coding schemes have been proposed for various transmission networks. One of them is the MPEG-4 fine granular scalability (FGS) scheme [1]. The FGS framework has a good balance between coding efficiency and scalability while maintaining a flexible and simple video coding structure. When compared with other error resilient streaming solutions, FGS has also demonstrated good error resilience attributes under packet losses. Moreover, FGS has recently been adopted by the MPEG-4 standard as the core coding method for video streaming applications.

Since the first version of the MPEG-4 FGS standard, there have been several improvements introduced to the FGS framework [2]. First, a very simple residual computation approach was proposed. This approach provides the same or better performance than the performance of more elaborate residual computation methods. Second, an adaptive quantization approach was proposed, and it results in two FGS-based video coding tools: frequency weighting and selective enhancement. Third, a hybrid-FGS scalability structure was also proposed. This structure enables us signal-to-noise ratio (SNR) scalable, temporal scalable, or both temporal-SNR scalable video coding and streaming [2].

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 159–166, 2004.

Figure 1 shows the encoder structure of the two-layer FGS system. In Fig. 1, the encoder estimates the channel capacity before encoding, and compresses the base layer using coding bits less than the channel capacity. Therefore, transmission of the base layer bitstream is always guaranteed. In the base layer, the main information of the input signal is coded in the same way as the traditional block-based coding scheme. In the enhancement layer, the residual data that is not coded in the base layer is divided into non-overlapping 8 × 8 blocks and each block is DCT transformed. All the 64 DCT coefficients in each block are zigzag-scanned and represented by binary numbers. These binary values form several bit-planes and entropy-coded to produce the output bitstream [1,3].

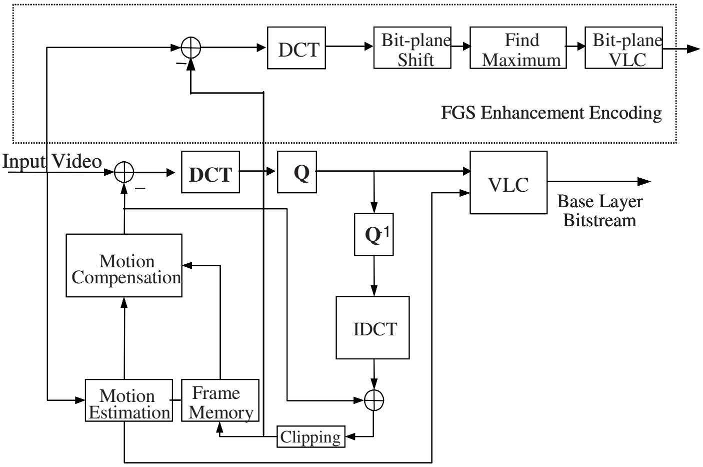


Fig. 1. FGS Encoder

Several advantages of FGS come at the expense of video quality reduction. FGS scarifies up to 2-3 dB in SNR, compared to nonscalable video coding scheme [4]. In order to overcome the performance degradation, we propose a new method of frequency weighting and selective enhancement for FGS. In the proposed frequency weighting method, we design a new frequency shifting matrix based on the human visual sensitivity function. In the selective enhancement method, the encoder decides visually important macroblocks (MB) automatically using the motion vector and position information of MB. We also define a new error metric to measure subjective image quality.

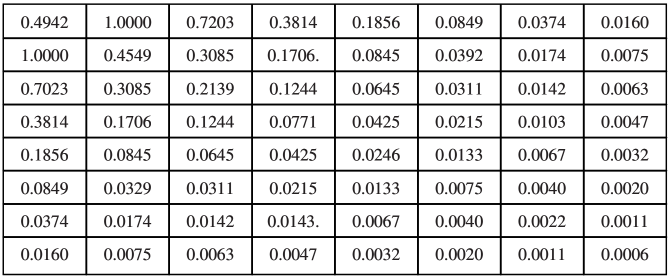
The paper is organized as follows. In Section 2, we describe a frequency weighting method based on the human visual system (HVS) and its implementation by bit-plane coding. In Section 3, we explain a fast selective enhancement method using coding information, such as the motion vector and the position information of each MB. In Section 4, we propose a new error metric to estimate the subjective image quality. After experimental results are presented in Section 5, we conclude this paper in Section 6.

2 HVS-Based Frequency Weighting

In general, human eyes are more sensitive to low frequency components than to high frequencies [5]. In order to improve visual quality of images, we can exploit the modulation transfer function (MTF) that represents the importance of each frequency component in terms of HVS. MTF can be described by

$$H(f) = a(b + cf)exp(-cf)^{d} \tag{1}$$

where f is the radial frequency in cycles/degree of the visual angle, and a, b, c and d are constants. Using the convolution-multiplication property of the DCT for a sampling density of 64 pels/degree, we can develop an 8×8 weighing matrix representing the HVS sensitivity [5][6]. Each 8×8 DCT coefficient is multiplied by the corresponding element of the frequency weighting matrix, reflecting their importance on HVS. Fig. 2 shows a typical frequency weighting matrix [5].



0.4942	1.0000	0.7203	0.3814	0.1856	0.0849	0.0374	0.0160
1.0000	0.4549	0.3085	0.1706.	0.0845	0.0392	0.0174	0.0075
0.7023	0.3085	0.2139	0.1244	0.0645	0.0311	0.0142	0.0063
0.3814	0.1706	0.1244	0.0771	0.0425	0.0215	0.0103	0.0047
0.1856	0.0845	0.0645	0.0425	0.0246	0.0133	0.0067	0.0032
0.0849	0.0329	0.0311	0.0215	0.0133	0.0075	0.0040	0.0020
0.0374	0.0174	0.0142	0.0143.	0.0067	0.0040	0.0022	0.0011
0.0160	0.0075	0.0063	0.0047	0.0032	0.0020	0.0011	0.0006

Fig. 2. Frequency Weighting Matrix

In order to provide HVS-based frequency weighting, we multiply each DCT coefficient by its corresponding element of the frequency weighting matrix. Therefore, the frequency weighted DCT coefficient is described by

$$C^{'}(i, j, k) = f_{w}(i) \cdot C(i, j, k) \tag{2}$$

where *C(i, j, k)* represents the DCT coefficient of the i-th component in the j-th block of the k-th MB, and *C'(i, j, k)* is the frequency weighted coefficient value by $fw(i)$ that is the frequency weight of the i-th DCT coefficient in each block.

We also convert the frequency weighting matrix to the frequency shift matrix. In order to make an appropriate mapping, we select the maximum shift factor *maxn(fw)* that represents the number of bits to be shifted up at the most important DCT coefficient. In the frequency weighting matrix, weighting values are normalized by one. Therefore, the frequency weighting matrix should be multiplied by $2^{maxn(fw)}$. After scaling the frequency weighting matrix, we transform it to the frequency shift matrix. As a result, the frequency shift matrix is obtained by

$$n_{fw(i)} = \lfloor \log_{2}^{\lceil} 2^{maxn(fw)} \cdot fw(i) \rceil \rfloor \tag{3}$$

3	4	3	2	1	0	0	0
4	3	2	1	1	0	0	0
3	2	2	1	1	0	0	0
2	2	1	1	0	0	0	0
1	1	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Fig. 3. Frequency Shift Matrix

where $n_{fw}(i)$ is a shift factor at the i-th DCT coefficient and $2^{maxn(fw)} \cdot fw(i)$ is the scaled frequency weighting. Figure 3 shows the frequency shift matrix with $maxn(fw)$=3.

Figure 4 represents the proposed frequency weighting process in the FGS enhancement layer, where we choose four for the maximum shift factor for the DC component.

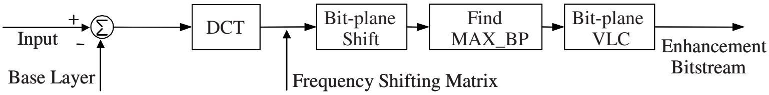


Fig. 4. Frequency Weighting for Enhancement Layer

3 Fast Selective Enhancement

In this section, we propose a fast *selective enhancement* (SE) algorithm using coding information, such as the motion vector and the location of MB, which can easily be extracted during the encoding process. Using this information, we can estimate the importance of each MB by

$$SE = P(x, y) \times ABS(mv_x) + ABX(mv_y) \tag{4}$$

where SE is the importance of the given MB, P(x ,y) is the position of the MB, $ABS(MV)$ is the absolute value of the motion vector. However, if we use only the coding information, we may miss some visually important MBs. Generally, if an MB is surrounded by visually important MBs, we can regard the MB as a visually important MB. Therefore, we apply lowpass filtering to SE values in each MB, as illustrated in Fig. 5

In Fig. 5, Vu, Hl, Hr, and Vd represent SE values of the surrounding MBs. Lowpass filtering is performed by

$$SE = (2SE + (Vu + Vd + Hl + Hr))/6 \tag{5}$$

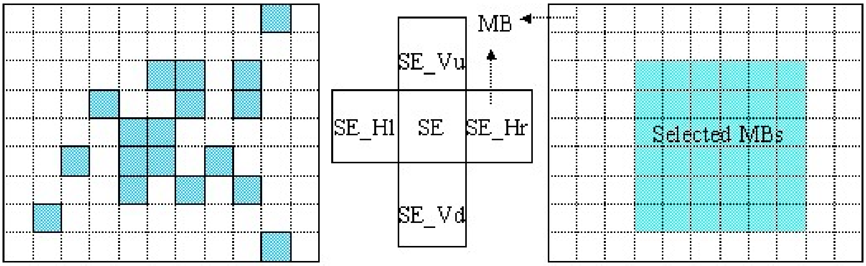


Fig. 5. Selective Enhancement Method

4 Perceptual Visual Quality

In this section, we define a new error metric to measure the subjective image quality based on the human visual system (HVS).

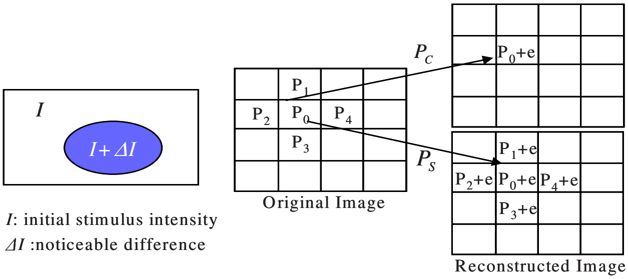


Fig. 6. Weber's Law and JNDE

According to the Weber's law, illustrated in Fig. 6, the minimum noticeable difference is proportional to the background intensity [7].

$$\frac{\Delta I}{I} = \alpha \tag{6}$$

In order to find the noticeable probability from the effect of the original pixel value, we change the Weber's law as follows [6]

$$\frac{\Delta I}{I} = \frac{D}{P} \geq \alpha \tag{7}$$

where p is the original pixel value and D is the difference between the original and its reconstructed values at a given pixel position. If the original image has a uniform distribution, the probability that the original pixel value is lower than the maximum threshold value p_{ths} is represented by

$$P_C = P(p \leq p_{ths}) = \frac{D/\alpha + 1}{2^n} \tag{8}$$

where n represents the number of bits assigned to each pixel. The noticeable probability P_S from the effect of the surrounding pixel values is

$$P_S = \sum_{k=1}^{4} k/4 \cdot_4 C_k(P_e)^k \cdot (1 - P_e)^{4-k} \tag{9}$$

where P_e represents the noticeable probability between the given error pixel and one of the neighboring pixels. *k/4* is the weighting factor for the number of k surrounding noticeable errors. As a result, the total noticeable probability P_{JNDE} of the given difference D is [6]

$$P_{JNDE} = P_C \cdot P_S \tag{10}$$

Until now, we introduce the just noticeable difference error (JNDE) using the Weber's law. We can also represent the peak signal-to-noise ratio (PSNR) by

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \tag{11}$$

where *MSE* represents the mean square error between the original and reconstructed images. In other words, *MSE* represents the average noise power. Applying the Parseval's theorem, *MSE* can be represented in the frequency domain.

$$MSE = \sum_{k=0}^{M-1} \sum_{v=0}^{N-1} \sum_{i=0}^{L-1} \alpha \cdot F_o(i, v, k) - F_r(i, v, k)^2 \tag{12}$$

where α represents a scaling factor between the frequency and spatial domains, and $F_o(i,\ v,\ k)$ and $F_r(i,\ v,\ k)$ represent DCT coefficients in the original and reconstructed images, respectively. In Eq. (12), M is the number of MBs in the frame, and N is the number of blocks in each MB. L denotes the number of pixels in the block. Consequently, we define a new error metric P_{HVS} by

$$P_{HVS} = \sum_{k=0}^{M-1} \sum_{v=0}^{N-1} \sum_{i=0}^{L-1} \alpha \cdot \frac{F_o^2(i, v, k)}{fw(i) \cdot \{F_o(i, v, k) - F_r(i, v, k)\}^2} \tag{13}$$

where $fw(i)$ the weighting factor obtained from the frequency weighing matrix in Fig. 2.

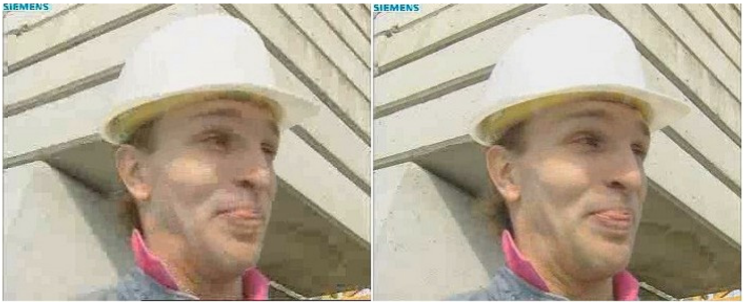
5 Experimental Results

In order to evaluate the performance of the proposed algorithm, we use the FOREMAN sequence, whose resolution is 352 × 288 pixels (CIF). Table 1 lists bit rates for the enhancement layers, where *FW0*, *FW1*, *FW2*, and *FW4* represent the maximum shift factor=0, 1, 2, and 4, respectively. Table 1 indicates that the frequency weighing method provides finer scalability than no frequency weighting method.

Table 1. Bit Rates for Enhancement Layers

Coded Bit-Plane	FW0	FW1	FW2	FW4
Base(kbit/s)	373	373	373	373
Base+E1	523	522	513	460
Base+E2+E2	1499	1164	1050	754
Base+E1+E2+E3	3645	2880	1914	1321
Base+E1+E2+E3+E4	7061	5696	3850	2312

Figure 7 shows the 6^{th} frame of the FOREMAN sequence. Fig. 7(a) is the reconstructed image with no frequency weighting, coded at 187.4 kbps. Fig. 7(b) is the reconstructed image with frequency weighting, coded at 165.2 kbps. From Fig. 7, we observe that perceptual quality of reconstructed images with frequency weighting is more acceptable than those without frequency weighting.



(a) (b)

Fig. 7. Comparison of Subjective Image Quality

Table 2. Number of Noticeable Errors

W	N	W-N	D	JND(W)	JND(N)	JND(W-N)
13,341	12,150	1,191	0	13,341	12,150	1,191
23,197	21,010	1,377	1	4,626	4,347	274
17,211	17,222	-11	2	6,790	6,795	-5
12,386	12,890	-504	3	7,306	7,603	-298
8,898	9,443	-545	4	6,986	7,414	-428
6,457	6,885	-428	5	6,330	6,751	-421
4,747	5,055	-300	6	4,747	5,055	-308
3,529	3,743	-214	7	3,529	3,743	-214

Table 2 lists the number of pixels at a given error (D) in both the frequency weighing case (W) and no frequency weighting case (N). We use α=0.02 to calculate the probability of noticeable error (JND (W, N)). JND (W) is obtained by multiply W with P_{JND}, which is calculated by Eq. (10). In the frequency weighting case, most errors are concentrated in the small error (D): therefore, we can obtain perceptually improved image quality in terms of HVS.

6 Conclusions

In this paper, we have proposed an HVS-based frequency weighting and a fast selective enhancement methods. In the proposed frequency weighting method, we assign frequency weighting to each DCT coefficient according to the human visual sensitivity function. We also convert the frequency weighting matrix to the frequency shift matrix to apply the frequency weighting method to the bit-plane coding. In the proposed selective enhancement method, we only use the coding information obtained in the encoding process. With the proposed ideas, we have obtained perceptually improved image quality. We have also defined a new error metric to measure perceptual visual quality of reconstructed images, both in the time and frequency domains.

Acknowledgements. This work was supported in part by Gwangju Institute of Science and Technology (GIST), in part by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC) at GIST, and in part by the Min-istry of Education (MOE) through the Brain Korea 21 (BK21) project.

References

1. Li, W.: Overview of Fine Granular Scalability in MPEG-4 Video Standard. IEEE Trans. on Circuit and System for Video Technology (2001) 301–317
2. Radah, H., Van der Schaar, M., and Chen, Y.: The MPEG-4 Fine Grained Scalable Video Coding Method for Multimedia Streaming over IP. IEEE Trans. Multimedia (2001) 53–68
3. Van der Schaar, M. and Radah, H.: A Hybrid Temporal SNR Fine Granular Scalability. IEEE Trans. Circuit and System for video Technology (2001) 318–331
4. Ling, F., Li, W., and Sun, H.: Bit-Plane Coding of DCT Coefficients for Image and Video Compression. Proc. SPIE, Visual Communication and Image Processing (1999) 25–27
5. Rao, K. and Yip, P.: Discrete Cosine Transform. Academic Press, New York, (1990)
6. Kim, S.H. and Ho, Y.S.: HVS-Based Frequency Weighting for Fine Granular Scalability. Proc. Information and Communication Technologies (2003) 127–131
7. Anil, K.J.: Fundamentals of Digital Image Processing. Prentice-Hall, (1989) 51

Efficient Multiview Video Coding Based on MPEG-4*

Wenxian Yang[1], King Ngi Ngan[2], and Jianfei Cai[1]

[1] School of Computer Engineering
Nanyang Technological University, Singapore
[2] Department of Electronic Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Abstract. In this paper, we propose an MPEG-4 based multiview video encoder. The main view of the sequences is encoded using the MPEG-4 encoder and the auxiliary views are encoded by joint motion and disparity compensation. The output of the encoder contains multiple bitstreams and the main bitstream can be decoded by a standard MPEG-4 decoder. Extensive experimental results show that our proposed multiview coding can achieve significant performance gain over the conventional multiview video encoder, which is implemented by applying the concept of the MPEG-2 Multi-View Profile (MVP) on the MPEG-4 platform. The improvements come from a more efficient reference structure and the joint estimation of disparity and motion fields in our proposed multiview coding system. In addition, in the case of five-view encoding, we also compare four different prediction structures in order to find the best structure under certain scenarios. The proposed encoder is very promising for the applications of video-conferencing and 3D telepresence.

1 Introduction

Researches in the fields of multimedia and virtual reality have stimulated increasing interest in 3D viewing. Three or more cameras may be used to form a multiocular system for the production of several image sequences obtained from slightly different viewpoints. The captured stereoscopic or multiview video sequences can provide more vivid information about the scene structure and a 3D feeling to the viewers through autostereoscopic display systems. Possible applications for multiview video include entertainment, education, medical surgery, communication, sightseeing and surveillance, etc [1]. However, since the amount of data increases dramatically with the number of views, compression is extremely important in the context of transmission and storage. Luckily, multiview video sequences can be efficiently compressed by exploiting the high correlations among different views, which is referred as disparity, in addition to the intra and inter frame redundancy within each view.

* This project is funded by the University Academic Research Fund (URC) RG3/02 - Depth-Based Video Segmentation of Nanyang Technological University, Singapore.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 167–174, 2004.

As indicated by Siegel [2], practical 3D-video compression methods typically consist of four modules: (1) coding one of the streams (the main stream) using a conventional method (e.g., MPEG), (2) calculating the disparity map(s) between corresponding points in the main stream and the auxiliary stream(s), (3) coding the disparity maps, and (4) coding the residuals. The MPEG-2 MVP is a straightforward way under such purpose. As discussed in [3], there are two prediction configurations for using MVP to encode the stereoscopic video sequences, where the configuration 2 has better results than the configuration 1. The average luminance PSNR values of the MVP configuration 2 can achieve up to 1.6 dB over the simulcast approach [4], in which each view is independently coded. Similarly, Luo [5] proposed MPEG-2 compatible stereoscopic video coders and adopted disparity compensation to remove the inter-channel redundancy. Almost all the existing stereoscopic and multiview encoders are based on MPEG-2. This is possibly because of the increasing demand on the applications of stereoscopic/multiview TV.

It is expected that MPEG-4 will play an important role in the area of multiview video coding especially for interactive applications, e.g., navigation through virtual 3D worlds with embedded natural video objects [5]. In our previous work [6], we have proposed a preliminary stereoscopic video encoder based on MPEG-4 and here we extend our work to encode multiple views with more extensive experimental results including the comparison with the MPEG-2 MVP and the comparison among different prediction structures.

This paper is organized as follows. The structure of the proposed system and the implementation details are described in Section 2. The simulation results are presented in Section 3 and finally the conclusions are given in Section 4.

2 Proposed Multiview Video Coding Scheme

2.1 Encoder Structure

New picture types are introduced for auxiliary views that use different prediction methods. For simplicity, we use subscript D to represent auxiliary view pictures corresponding to the main view picture type, to indicate that this is

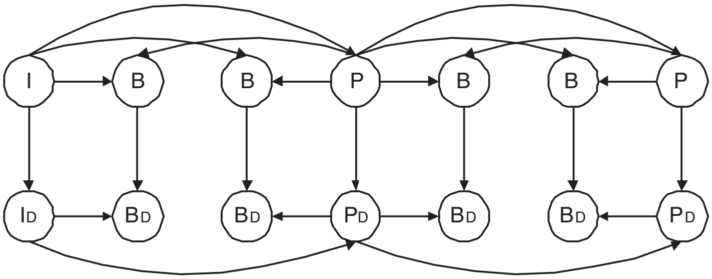


Fig. 1. The GOP structure

also predicted by disparity. Figure 1 shows a simple GOP structure for stereoscopic video sequences that is extended from MPEG. In our proposed encoder, users can define the GOP structure by setting the M and N parameters. Here, N is intra distance, which is the length of a GOP, and M is the prediction distance. The frame structure of I, P and B frames in MPEG provides the functionalities of random access, editability and independently decodability of video segments [2]. As shown in Figure 2, we retain the frame structure for the main view and introduce new picture types I_D, P_D and B_D for the auxiliary view, where I_D VOPs are predicted by disparity and P_D/B_D VOPs are predicted jointly by disparity and motion fields.

Disparity vectors are encoded by DPCM and Huffman coding, which is similar to the coding of motion vectors as in MPEG-4. Residual data after disparity/motion compensation is encoded by the block-based DCT coding as defined in MPEG-4.

2.2 Reference Structure

View level reference structure refers to the problem of given a (n, m) pair, where n is the total number of views and m is the number of main views, how to determine the position of the main views. In our case, we only consider the situation that all the views are parallel with equal distance between the adjacent pairs. The basic principle for selecting the reference structure is based on the fact that the further the distance between the two views is, the more occlusion occurs and the worse the disparity estimation will be. For 2-view and 3-view video coding, the configuration is quite straightforward. In our simulations, the left view is coded as the main view and the right view is coded as the auxiliary view for 2-view video. For 3-view video, the middle view is selected as the main view to provide better prediction to the two auxiliary views.

For the 5-view video, we consider the four configurations as shown in Figure 2. The bold lines represent the main views and the others are the auxiliary views predicted from the main views. Config 1 adopts directly the concept from the GOP structure in mono-view video coding. In particular, in Config 1, View 4 is predicted from view 0, and the other views are predicted bi-directionally from

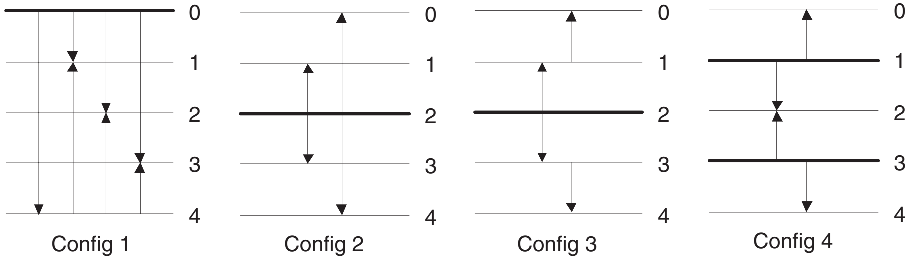


Fig. 2. View-level prediction structures for 5-view video encoding

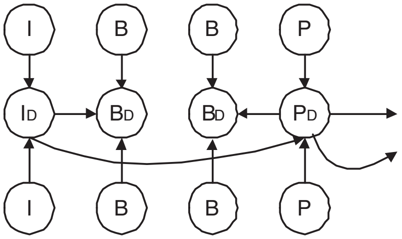


Fig. 3. Picture-level prediction structures for 5-view video coding

View 0 and View 4. In this way, high compression can be achieved especially when the disparity between the views is very small. In both Config 2 and Config 3, the middle view is coded as main view and the difference is whether the outmost views are predicted from the main view or the adjacent auxiliary views. In Config 4, there are two main views and this is suitable for the case when the disparity between each adjacent view pairs is very large.

Picture level reference structure considers the selection between disparity and motion vectors, forward motion and backward motion vector, etc. The selection is done at macroblock level based on the minimum distortion criterion. As shown in Figure 1, the auxiliary view pictures are predicted either by disparity alone (I_D) or jointly by disparity and motion fields (P_D and B_D). In Config 1 and Config 3 there's another kind of auxiliary view which is predicted by two main views as shown in Figure 3. In this case the I_D pictures are predicted by bi-directional disparity and the P_D/B_D pictures are predicted by two disparity fields and one motion field.

3 Simulation Results and Analysis

3.1 Video Capture and Preprocessing

We captured 3-view (left, middle and right views) video data, Reading, using our multi-camera system. The system is composed of three cameras, each controlled by a software motion controller. The three cameras are parallel placed on a horizontal line with an equal distance of 62 mm between two adjacent ones. For the Reading sequences, the focus length for the three cameras is 16 mm. The approximate object distance is 3 meters. The frame rate is 30 frames per second and the resolution is 768 by 576 pixels. In this work, we only test the even field of the obtained sequences, i.e., 768 by 288 pixels. The Reading sequence has static and homogeneous background, small motion for the head and the body, medium motion for the hands. The disparity range of Reading is very large and the corresponding search range is set to $[-64, 63]$ pixels, while the motion is relatively small and the corresponding search range is $[-8, 7]$ pixels. The standard Train & Tunnel stereoscopic video sequences are also used for testing our system. The Train & Tunnel sequence is 720 by 576 pixels at the frame rate of 25 fps. The search range for Train & Tunnel is $[-8, 7]$ pixels for both disparity and motion fields. The first frames of the original video sequences are shown in Figure 4. For testing the 5-view video coding for Reading and the 3/5-view video coding for Train & Tunnel, we generate intermediate views using the IVR algorithm [7].

The auxiliary video sequences are input to the encoder after balancing with the main video sequence to eliminate the potential signal difference caused by light conditions and camera differences.

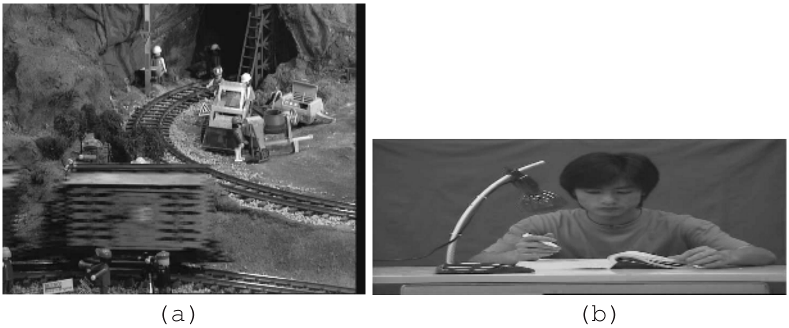


Fig. 4. Original video sequences, the first frame of (a) left view of Train & Tunnel and (b) middle view of Reading.

3.2 Encoding 2-View and 3-View Video Sequences

Without loss of generality, we set the intra distance $N = 9$ frames and the prediction distance $M = 3$ frames. Only luminance information is encoded for all the simulations. We extend the MPEG-4 rate control scheme as defined in MPEG-4 VM 18.0 [8] for our multiview video encoder. Three more picture types (I_D, P_D and B_D) are introduced and all the views are controlled using one common buffer.

As mentioned in the introduction section, the configuration 2 of MPEG-2 MVP improves the PSNR up to 1.6 dB than the simulcast scheme. For fair comparison, we implement the MPEG-2 MVP configuration 2 on the MPEG-4 platform. Thus, the major difference between our proposed system and the conventional scheme is the prediction structure of the auxiliary view and the estimation of disparity and motion fields for both main view and auxiliary views. In the MVP configuration 2, the auxiliary view pictures are predicted by disparity from its corresponding main view picture and by motion from its most adjacent auxiliary view picture. The disparity and motion vectors are searched independently using the full search block matching method.

The coding results for 2-view video encoding are shown in Table 1 and the results for encoding 3-view video are shown in Table 2. The total BW is the

Table 1. The comparison of the 2-view encoding results.

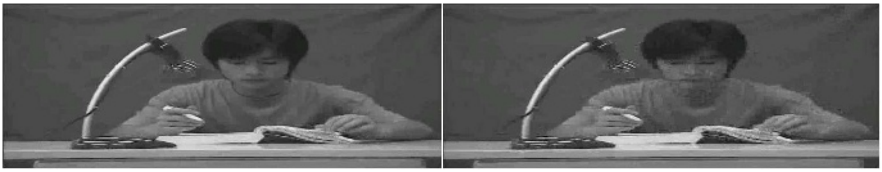
Sequences	Total BW	Proposed		Conventional	
	(Mbps)	Auxi BR(%)	Overall PSNR	Auxi BR(%)	Overall PSNR
Train & Tunnel	2	40.29	28.94	45.32	28.03
	4	40.67	31.91	46.56	30.89
	6	40.11	33.90	44.42	33.02
	8	44.02	35.22	47.06	34.24
Reading	1	47.52	34.22	51.80	32.45
	1.5	49.30	37.38	51.93	36.54
	2	49.02	39.35	53.56	37.70
	2.5	50.22	40.88	54.06	39.70

Table 2. The comparison of the 3-view encoding results.

Sequence	Total BW	Proposed			Conventional		
	(Mbps)	Auxi1 BR(%)	Auxi2 BR(%)	Overall PSNR	Auxi1 BR(%)	Auxi2 BR(%)	Overall PSNR
Train & Tunnel	3	25.12	27.34	29.79	25.57	26.69	28.89
	6	26.81	30.73	32.39	29.42	32.47	31.34
	9	27.42	30.27	34.03	29.92	33.59	33.06
	12	28.71	30.42	35.29	29.79	30.00	34.54
Reading	1.5	34.88	31.53	34.87	39.12	33.55	32.36
	2	36.70	30.89	36.54	37.54	33.09	34.99
	3	35.98	33.33	39.59	38.96	32.92	37.52
	4	34.86	32.21	40.42	38.80	34.07	39.35

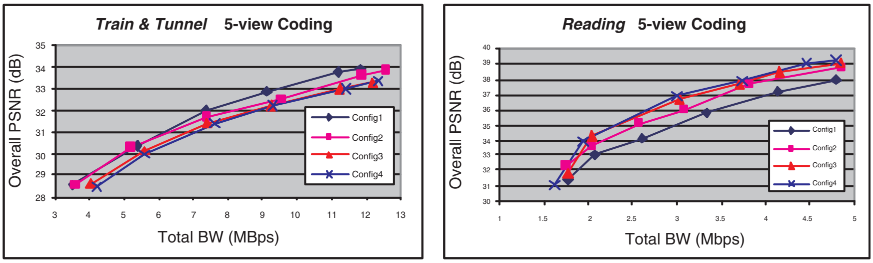
total bandwidth for all the views and the overall PSNR is obtained from the average distortion (MSE) of all the pictures in all the views. The Auxi BR (%) indicates the percentage of bits used by the auxiliary view. As shown in Table 1, the proposed encoder increases the overall PSNR about 1 dB for Train & Tunnel and 1 to 1.6 dB for Reading. For encoding three views, the PSNR gain is around 1 to 2 dB for Reading. It is interesting to see that the quality of the auxiliary view of our proposed encoder is always better than that of the conventional encoder although our proposed encoder uses lesser bits for encoding the auxiliary views. One reason is that the prediction frames from the main view in our proposed encoder has higher quality. Moreover, in our proposed system, disparity and motion fields for the auxiliary views are obtained from joint regularization but not by full search, which reduce the bits for coding the vectors for P_D/B_D pictures. Since more bits are left for encoding the main view, the main view quality of our proposed encoder is always better, as shown in Table 1 and Table 2.

To compare the coding results subjectively, we also show one frame of the decoded images obtained by MPEG-4 and our proposed encoder for Reading using the same encoding parameters in Figure 5. From these results we can see that our proposed system is efficient for encoding the multiview video data.



(a) (b)

Fig. 5. Decoded auxiliary view image for Reading (a) Proposed scheme and (b) Conventional scheme.



(a) (b)

Fig. 6. Comparison of view-level prediction structures for 5-view video coding, (a) Train & Tunnel and (b) Reading.

3.3 Comparison of Different Prediction Structures for 5-View Encoding

The comparison of the four configurations in view-level prediction is shown in Figure 6 (a) and (b) for Reading and Train & Tunnel, respectively. For Reading, since the disparity range is very large, Config 4 obtains the best result, where View 1 and View 3 are encoded as main views. Config 1 has the worst result since the direct disparity between the two outmost views (View 0 and View 4) is very large and thus disparity prediction for View 4 from View 0 fails. For Train & Tunnel, since the disparity range is small, Config 1 has the best result, which is more obvious at higher bitrates. Config 4 has the worst result since two main views cost a lot of bits. For Config 3, the results are unstable and largely depend on the rate control algorithm. This is because, in Config 3, View 1 and View 3 are predicted from the main view, View 2, but they are also the reference views for View 0 and View 4 respectively. The reconstructed image qualities of View 1 and View 3 have great effects on the image qualities of View 0 and View 4.

From these simulation results, we confirm that the best view-level prediction structure very much depends on the properties of the multiview video data, i.e., the disparity range. For a small baseline distance in the applications such as robotic stereovision, Config 1 will be a good choice. However, for a large baseline distance, as in videoconferencing, Config 4 can achieve better performance. However, in order to fairly evaluate the different prediction structure, the rate control algorithm needs to be well designed and finely tuned. This is under investigation.

4 Conclusions

An MPEG-4 based multiview video encoder has been proposed in this paper, which well exploits not only the redundancy within each view but also among different views. The extensive experimental results have demonstrated that our proposed encoder significantly outperforms the conventional scheme based on the concept of MPEG-2 MVP. We have also compared different prediction structures in the case of 5-view coding and have found that the optimal prediction structure actually depends on the video properties, i.e., camera parameters, disparity range and the applications.

The encoder can be easily extended to encode a various numbers of views, and can also be extended to handle multiple video objects, given the original shape masks of the sequences. Future work will focus on rate control for multiview video encoder and also consider the HVS (human visual system) factor.

References

1. A.Smolic and H.Kimata.: Applications and requirements for 3DAV. ISO/IEC JTC1/SC29/WG11 N5877. (July 2003)
2. M. W. Siegel, S. Sethuraman, J. S. McVeigh and A. G. Jordan.: Compression and interpolation of 3D-stereoscopic and multi-view video. Stereoscopic Displays and Virtual Reality Systems IV, Proceedings of the SPIE. vol. 3012 (Feb. 1997) 227–238
3. A.Puri, R.V.Kollarits and B.G.Haskell.: Stereoscopic video compression using temporal scalability. Visual Communication: Image Processing. vol. 2501 (1995) 745–756
4. Xuemin Chen and Ajay Luthra.: MPEG-2 multi-view profile and its application in 3DTV. SPIE/IS&T Multimedia Hardware Architectures. (Feb. 1997) 212–223
5. Luo Yan, Zhang Zhaoyang and An Ping.: Stereo video coding based on frame estimation and interpolation. IEEE Trans. on Broadcasting. vol. 49. no. 1 (March 2003) 14–21
6. W.Yang and K.N.Ngan.: MPEG-4 based stereoscopic video sequences encoder. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (May 2004)
7. H.S.Kim and K.H.Sohn.: Feature-based disparity estimation for intermediate view reconstruction of multiview images. International Conference on Imaging Science, Systems, and Technology. vol. 2 (June 2001) 1–8
8. Weiping Li, J-R Ohm, Mihaela van der Schaar, Hong Jiang and Shipeng Li.: MPEG-4 video verification model version 18.0. ISO/IEC JTC1/SC29/WG11 N3908. (Jan. 2001)

Block Matching Using Integral Frame Attributes

Viet Anh Nguyen and Yap-Peng Tan

School of Electrical and Electronic Engineering, Nanyang Technological University,
50 Nanyang Avenue, Singapore 639798, Singapore

Abstract. Block-based motion estimation is widely used in video compression for exploitation of video temporal redundancy. Although effective, the process is arguably the most computationally intensive step in a typical video encoder. To speed up the process, a large number of fast block-matching algorithms (BMAs) have been proposed for motion estimation by limiting the number of search locations or simplifying the measure of match between the two blocks under comparison. In this paper we propose a new BMA that measures the match by using such block features as block mean and block variance, quantities that can be easily computed from integral frame attributes. Experimental results show that the proposed BMA can reduce notably the computational load and achieve compression performance very close to that of existing BMAs using conventional block-matching measures.

1 Introduction

Motion-compensated prediction (MCP) is a popular scheme used in video compression for exploiting the temporal redundancy between neighboring video frames. It achieves compression by representing each frame with motion vectors compensating its scene change with respect to a reference frame. Owing to its ease of computation and hardware implementation, block-based motion estimation has attracted considerable attention and been adopted in popular video compression standards, such as H.261/3/4 and MPEG-1/2/4, where some block-matching algorithm (BMA) is used to obtain the motion vector for each block. The most widely used BMA is probably the full search (FS) algorithm, which estimates the motion vector for each block by matching it with all the candidate blocks within a search window in a reference frame. Being exhaustive, the FS algorithm achieves optimal performance, but its high computational requirement makes it unsuitable for many real-time applications.

To reduce the computational complexity, many fast BMAs have been developed to obtain sub-optimal matches by limiting the number of search locations in the reference frame. These include the three-step search (TSS) [1], new three-step search (NTSS) [2], 2-D logarithm search (2-DLOG) [1], four-step search (FSS), and cross-diamond search (CDS) [3]. These BMAs assume the match of two blocks, gauged by some block-matching (BM) measure, decreases monotonically as the search location moves away from the best match (i.e., the one with the optimal BM measure). As such assumption may not hold for all cases, many of these existing BMAs share a common problem: being trapped in a local optimum and thus degrading compression performance.

Besides limiting the number of search locations, another means to reduce the computational cost of a BMA is by simplifying the BM measure evaluated at each search

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 175–183, 2004.

location. Koga et al. [4] suggest a pixel decimation scheme for computing the BM measure based on a set of pixel patterns; they define some representative pixel patterns and select the most suitable ones based on the content of each block. In [5], horizontal or vertical projections of pixel values in blocks are used to compute the BM measure. The work shares some similarity with our proposed method. However, as the computations of these projection values and their derived BM measures are still quite intensive, the existing projection methods can only reduce the computational complexity by a factor equal to half of the block size (i.e., a factor of 8 for the typical 16 × 16 blocks). In comparison, the method proposed in this paper can reduce the complexity by factors of 21-36 for a typical search window size, as we shall show later.

In this paper, we propose a fast BMA that makes use of such block features as block mean and block variance to compute the BM measure. Specifically, each block under consideration is partitioned into a number of sub-blocks arranged in a fixed pattern, and the mean or variance of intensity in each sub-block is matched with that of each candidate block. Inspired by the work of Viola et al. [6], we first compute the integral attributes of each frame to allow for very fast computation of the proposed BM measures. In conjunction with the FS or a fast search algorithm, our proposed BM measures can achieve compression performance close to that using the existing sum-of-absolute-differences (SAD) BM measure, but incurring a much lower computational cost.

2 Integral Frame Attributes

Given a video frame, let $g(m,n)$ be a frame attribute characterizing some measure of frame features about pixel (m,n). The integral frame attribute at pixel (m,n), denoted as $\mathcal{I}_g(m,n)$, is defined as the sum of the frame attributes $g(m,m)$'s over the region that is above and to the left of pixel (m,n), inclusive [6]; that is (see Fig. 1 for illustration)

$$\mathcal{I}_g(m,n) = \sum_{x=0}^{m}\sum_{y=0}^{n} g(x,y). \tag{1}$$

Let $\mathcal{R}_g(m,n)$ denote the cumulative row sum of frame attributes $g(m,n)$'s, defined as

$$\mathcal{R}_g(m,n) = \sum_{x=0}^{m} g(x,n). \tag{2}$$

Defining $\mathcal{R}_g(-1,n) = 0$ and $\mathcal{I}_g(m,-1) = 0$, one can compute the integral frame attribute $\mathcal{I}_g$ in one pass by using two recursive formulas:

$$\mathcal{R}_g(m,n) = \mathcal{R}_g(m-1,n) + g(m,n); \quad \mathcal{I}_g(m,n) = \mathcal{I}_g(m,n-1) + \mathcal{R}_g(m,n). \tag{3}$$

Hence, for a frame with $M \times N$ pixels, only $2MN$ additions are required to compute the integral attribute, excluding the computational cost for each frame attribute $g(m,n)$'s.

Using this integral frame attribute, the sum of the frame attributes in any rectangular block (hereafter referred to as block sum ($\mathcal{BS}_g$) for simplicity) can be computed with 3 arithmetic operations (1 addition and 2 subtractions). This can be seen from Fig. 2(a),

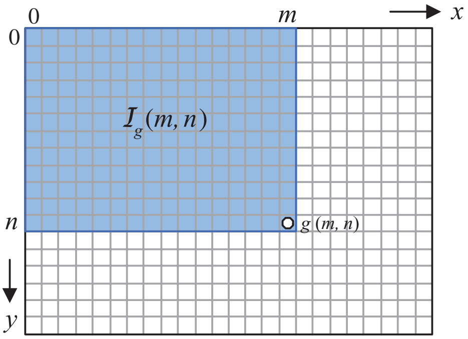


Fig. 1. The value of integral frame attribute at pixel (m, n) is equal to the sum of some frame features over the region that is above and to the left of pixel (m, n), inclusive, in the original frame.

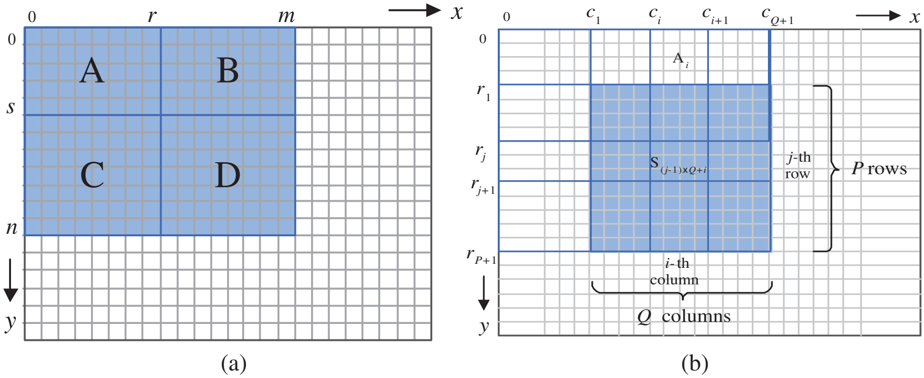


Fig. 2. (a) The sum of all frame features in block D can be computed by using the four corresponding integral frame attributes at the block boundaries. (b) It takes only $(2P+1)\times Q$ arithmetic operations to compute the proposed SAD-BM measure for a K-block pattern.

where the $\mathcal{BS}_g$ of block D (with support $\Omega_D = \{(x, y) : r < x \leq m, s < x \leq n\}$) can be computed by using four corresponding integral attributes at the block boundaries as

$$\mathcal{BS}_g(\mathrm{D}) = \sum_{x=r+1}^{m} \sum_{y=s+1}^{n} g(x, y) = \mathcal{I}_g(m, n) - \mathcal{I}_g(r, n) - \mathcal{I}_g(m, s) + \mathcal{I}_g(r, s) \quad (4)$$

Let $\mathcal{I}_f$ and $\mathcal{I}_{f^2}$ denote the integral frame attributes when $g(m, n)$ is the gray-level of pixel (m, n), denoted as $f(m, n)$, and the square of gray-level value at pixel (m, n), respectively. Then the block sums of pixel values and the square of pixel values, denoted as $\mathcal{BS}_f$ and $\mathcal{BS}_{f^2}$ respectively, can be computed using (4) with $\mathcal{I}_f$ and $\mathcal{I}_{f^2}$. Note that one multiplication is required to compute the square of a pixel's gray-level value. If one multiplication needs three arithmetic operations [7], a total of $5MN$ arithmetic operations are required to compute the integral frame attribute $\mathcal{I}_{f^2}$.

By using integral frame attribute, the block mean ($\mathcal{BM}$), which is regarded as a potential block feature, can be obtained by

$$\mathcal{BM}(\mathrm{D}) = \mathcal{BS}_f(\mathrm{D})/N_D, \tag{5}$$

where N_D is the total number of pixels in block D. In addition, we can obtain the variance of all pixel values in a rectangular block D through its $\mathcal{BS}_f$ and $\mathcal{BS}_{f^2}$ by

$$\delta^2(\mathrm{D}) = \frac{1}{N_D}\mathcal{BS}_{f^2}(\mathrm{D}) - \left[\frac{1}{N_D}\mathcal{BS}_f(\mathrm{D})\right]^2 \tag{6}$$

3 Proposed Block-Matching Algorithm

Widely used as a BM measure in many existing BMA's is the sum of absolute differences of pixel values (SAD) between the current block and the candidate block, defined as

$$\mathrm{SAD} = \sum_{x=1}^{N_1}\sum_{y=1}^{N_2} |f_c(x,y) - f_r(x,y)|, \tag{7}$$

where $f_c(x,y)$ and $f_r(x,y)$ denote the pixel values from the current and the candidate block of $N_1 \times N_2$ pixels, respectively.

The SAD measure generally provides good matching precision, but it is computationally intensive. To minimize the number of computations required for motion estimation, we propose in this paper to make use of the integral frame attributes to perform the block matching. Specifically, each block under consideration is first partitioned into a number of sub-blocks arranged in a fixed pattern, and the integral frame attributes are then used to compute the feature of each sub-block as a BM parameter. The sub-block patterns that we have examined are shown in Fig. 3.

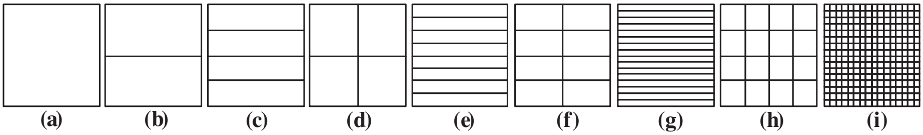


Fig. 3. Example block patterns: a) 1-block, (b) 2-block, (c) 4-strip, (d) 4-block, (e) 8-strip, (f) 8-rectangle, (g) 16-strip, (h) 16-block, and (i) SAD.

In the following, we examine three potential BM measures by using various block features and making use of the integral frame attributes.

(1) SAD of Block Means (SAD-BM): This measure uses the $\mathcal{BM}$ as a BM parameter. To locate the best-matching block in the reference frame, the $\mathcal{BM}$'s of all sub-blocks in the current block are compared with those in each candidate block. Specifically, the sum of absolute differences between the corresponding $\mathcal{BM}$'s is computed as the BM measure, given by

$$\text{SAD-BM} = \sum_k |\mathcal{BM}_c(\mathrm{S}_k) - \mathcal{BM}_r(\mathrm{S}_k)|, \tag{8}$$

where $\mathcal{BM}_c(\mathrm{S}_k)$ and $\mathcal{BM}_r(\mathrm{S}_k)$ denote the block means of the kth sub-blocks from the current and the candidate blocks, respectively, and can be computed by using (5).

(2) SAD of Block Variances (SAD-VR): This measure computes the sum of absolute variance differences between the corresponding sub-blocks as the BM measure, given by

$$\text{SAD-VR} = \sum_k \left|\delta_c^2(\mathrm{S}_k) - \delta_r^2(\mathrm{S}_k)\right|, \tag{9}$$

where $\delta_c^2(\mathrm{S}_k)$ and $\delta_r^2(\mathrm{S}_k)$ denote the variances of the kth sub-blocks from the current and the candidate blocks, respectively, and can be computed by using (6).

(3) SAD of Block Means and Variances (SAD-MV): This is a hybrid measure combining the above two BM measures. It consists of two parts, the sum of absolute block mean differences and the sum of absolute variance differences between the corresponding sub-blocks, defined as

$$\text{SAD-MV} = \sum_k |\mathcal{BM}_c(\mathrm{S}_k) - \mathcal{BM}_r(\mathrm{S}_k)| + \lambda \sum_k \left|\delta_c^2(\mathrm{S}_k) - \delta_r^2(\mathrm{S}_k)\right| \tag{10}$$

where λ is a proper weighting factor. By verification with a large number of test sequences using different values of λ, we note that when λ is set to 0.02, the proposed SAD-MV measure generally provides the best compression performance. Hence, in our implementation of the proposed SAD-MV measure, we set λ to 0.02.

4 Computational Gains

Consider a video frame of $M \times N$ pixels, a block size of 16×16 pixels, and a search window size of $\pm W$ pixels for block-matching motion estimation. Assume a block pattern of K equal sub-blocks, partitioned into P rows and Q columns, to be used for evaluating the BM measure (see Fig. 2(b) for illustration). To compute the proposed SAD-BM measure, we can calculate the K $\mathcal{BS}_f$'s in the current block and in each of its candidate blocks by exploiting the adjacent property of the sub-blocks as follows. For each column i ($1 \le i \le Q$), we first obtain $\mathcal{BS}_f$'s of the following blocks from integral frame attribute $\mathcal{I}_f$ with $P + 1$ subtractions:

$$\mathcal{BS}_f(\mathrm{A}_i) = \mathcal{I}_f(c_{i+1}, r_1) - \mathcal{I}_f(c_i, r_1) \tag{11}$$

$$\mathcal{BS}_f(\mathrm{A}_i) + \sum_{j=1}^{t} \mathcal{BS}_f(\mathrm{S}_{(j-1)\times Q+i}) = \mathcal{I}_f(c_{i+1}, r_{t+1}) - \mathcal{I}_f(c_i, r_{t+1}), \tag{12}$$

for $1 \le t \le P$. The $\mathcal{BS}_f$'s of the P sub-blocks in the ith column can be computed from the above $(P+1)$ $\mathcal{BS}_f$'s by another P subtractions. Hence, to compute all $\mathcal{BS}_f$'s of the block under comparison, we need a total of $(2P + 1) \times Q$ subtractions.

In addition, to match the current block and a candidate block, $3 \times K - 1$ arithmetic operations (K subtractions, $K - 1$ additions, and K absolute conversions) are required to compute the sum of absolute differences for K corresponding pairs of $\mathcal{BS}_f$'s. Thus, the number of arithmetic operations required to compute the SAD-BM measure at each search location is equal to

$$\mathcal{C}_{\text{BM}} = 5 \times PQ + Q - 1. \tag{13}$$

Similarly, by using (6), the number of arithmetic operations required to compute the SAD-VR and SAD-MV measures at each search location can be given by, respectively,

$$\begin{aligned}\mathcal{C}_{\mathrm{VR}} &= 14 \times PQ + 2 \times Q - 1\\ \mathcal{C}_{\mathrm{MV}} &= 18 \times PQ + 2 \times Q - 2.\end{aligned} \tag{14}$$

Table 1 lists the numbers of arithmetic operations required to compute the proposed BM measures at each search location by using different block patterns.

Table 1. Numbers of arithmetic operations required to compute the proposed BM measures at each search location using different block patterns.

	Block partition		No. of operations		
Pattern	P	Q	$\mathcal{C}_{\mathrm{BM}}$	$\mathcal{C}_{\mathrm{VR}}$	$\mathcal{C}_{\mathrm{MV}}$
1-block	1	1	5	15	18
2-block	2	1	10	29	36
4-strip	4	1	20	57	72
4-block	2	2	21	59	74
8-strip	8	1	40	113	144
8-rect	4	2	41	115	146
16-strip	16	1	80	225	288
16-block	4	4	83	231	294

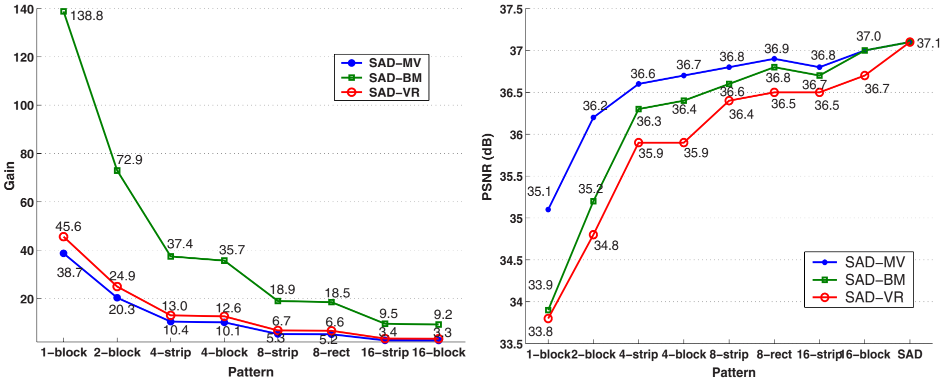
It should be noted that when block-matching is performed directly on the input video sequence, another $2MN$ and $5MN$ arithmetic operations (see Eq. (3)) are required to compute each integral frame attribute $\mathcal{I}_f$ and $\mathcal{I}_{f^2}$, respectively; moreover, the sub-block features computed for each current frame can be reused when the frame becomes a reference frame later.

On the other hand, when the existing SAD measure is used to match two blocks, at each search location 16×16 pixel pairs are to be compared, and each comparison requires 3 operations—a subtraction, an addition, and an absolute conversion.

Table 2 shows the total number of operations required per frame when each of the proposed BM measures or the existing SAD measure is used for motion estimation in conjunction with the popular FS and TSS algorithms. For a search window with size of $\pm W$ pixels, the number of locations to be searched by the FS algorithm for each block is equal to $(2W+1)^2$, and that by the TSS algorithm is $1 + 8 \times \lceil \log_2 W \rceil$, where $\lceil\cdot\rceil$ denotes the ceil operator. Hence, in comparison to the SAD measure, the number of the arithmetic operations required by the proposed BM measures can be reduced by factors of $3 \times (2W+1)^2/(\mathcal{R}+\mathcal{C}\times(2W+1)^2/16^2)$ and $3 \times (1 + 8 \times \lceil \log_2 W \rceil)/(\mathcal{R}+ \mathcal{C} \times (1 + 8 \times \lceil \log_2 W \rceil)/16^2)$ using the FS and TSS algorithms, respectively, where $\mathcal{C}$ is the number of operations required to compute the proposed BM measures at each search location by using different block patterns (as listed in Table 1), while constant $\mathcal{R}$ is equal to 2 for the SAD-BM measure and equal to 7 for either SAD-VR or SAD-MV measure, respectively. Fig. 4(a) plots the gain obtained by different proposed BM measures compared with the conventional SAD measure using the FS algorithm.

Table 2. Numbers of arithmetic operations required by the proposed BM measures and the conventional SAD measure when used together with the FS and TSS algorithms, respectively.

BMA	BM measure	No. of arithmetic operations per frame
FS	SAD	$3MN \times (2W+1)^2$
	SAD-BM	$2MN + \mathcal{C}_{\mathrm{BM}} \times (2W+1)^2 \times MN/16^2$
	SAD-VR	$7MN + \mathcal{C}_{\mathrm{VR}} \times (2W+1)^2 \times MN/16^2$
	SAD-MV	$7MN + \mathcal{C}_{\mathrm{MV}} \times (2W+1)^2 \times MN/16^2$
TSS	SAD	$3MN \times (1 + 8 \times \lceil \log_2 W \rceil)$
	SAD-BM	$2MN + \mathcal{C}_{\mathrm{BM}} \times (1 + 8 \times \lceil \log_2 W \rceil) \times MN/16^2$
	SAD-VR	$2MN + \mathcal{C}_{\mathrm{VR}} \times (1 + 8 \times \lceil \log_2 W \rceil) \times MN/16^2$
	SAD-MV	$2MN + \mathcal{C}_{\mathrm{MV}} \times (1 + 8 \times \lceil \log_2 W \rceil) \times MN/16^2$



(a) (b)

Fig. 4. (a) The gain obtained by different proposed BM measures in comparison with the FS algorithm using the conventional SAD measure. (b) Performance comparison obtained by using the proposed BM measures computed with different block patterns for the Foreman sequence.

5 Experimental Results

We have conducted a series of experiments to evaluate the performance of the proposed BMA. Our test sequences include ten popular CIF resolution (352×288) sequences, as shown in Table 3. These sequences contain different amounts of motion and spatial details, and have been widely tested in the research of video compression.

We conducted the experiments by using the Test Model 5 (TM5) MPEG-2 encoder provided by the MPEG Software Simulation Group at MPEG.org [8]. For each test sequence, we set the target bit-rate, frame rate and search window size to 1.5 Mbits/s, 30 frames/s and $W = 15$, respectively.

The first set of experiments was conducted to evaluate the compression performance of different proposed BM measures by using various block patterns as shown in Fig. 3. Fig. 4(b) shows the average PSNR results for the Foreman sequence obtained by using the three proposed BM measures in conjunction with the FS algorithm. The results show that, in comparison with the existing SAD measure, our BM measures can re-

Table 3. PSNR results (in dB) of ten test sequences obtained by the FS and TSS algorithms using the proposed BM measures with the 4-block pattern and the existing SAD measure

	FS				TSS			
Sequence	SAD	BM	VR	MV	SAD	BM	VR	MV
Coastguard	34.5	33.6	32.9	34.0	34.3	33.4	32.7	33.9
Container Ship	38.8	38.6	37.9	38.8	38.8	38.7	38.4	38.8
Flower Garden	29.4	28.7	27.8	29.1	28.8	27.8	27.0	28.5
Football	29.5	28.9	28.4	29.2	29.2	28.8	28.4	29.0
Foreman	37.1	36.4	35.9	36.7	36.6	36.1	35.7	36.3
M & D	41.1	41.0	40.7	41.0	41.1	41.0	40.8	41.0
News	41.5	41.4	41.2	41.4	41.5	41.3	41.2	41.4
Stefan	31.8	30.9	30.1	31.4	30.9	30.3	29.6	30.9
Tempete	31.5	31.0	30.4	31.3	31.5	31.1	30.8	31.3
Tennis	30.6	30.0	29.5	30.2	30.3	30.1	29.6	30.2

duce a significant computation load in motion estimation without degrading much the compression performance. It is noted that, although more computationally intensive, the SAD-VR measure cannot provide a better compression performance as compared with the SAD-BM measure. However, when block variances are combined with block means with a proper weighting factor, the SAD-MV measure is able to perform better than SAD-BM. The computational complexities of the SAD-VR and SAD-MV measures are much higher than that of SAD-BM measure.

In addition, the performance improves as the number of the sub-blocks increases (i.e., the size of each sub-block decreases), for the reason that the video spatial variations can be gauged more closely when block-matching is evaluated based on means and variances in smaller sub-blocks. However, it is also evident that only a marginal performance gain can be obtained by using block patterns with sizes of sub-blocks smaller than that of the 4-strip or 4-block pattern. Particularly, the performance of the 4-strip or 4-block pattern is only a little worse than that of the 8-strip or 8-rect pattern, whereas the complexity is substantially better. In other words, the 4-strip and 4-block pattern can provide a good tradeoff between the complexity and compression performance.

In another set of experiments, we encoded the ten test sequences using the proposed BM measures with the 4-block pattern and the existing SAD measure in conjunction with the FS and TSS algorithms. The group-of-pictures (GOP) of each encoded sequence consists of one intra-coded frame followed by nine predictive-coded frames. For comparison, the PSNR results of the proposed BM measures are provided in Table 3. The results show that, incurring a much lower computational cost, the proposed BM measures can perform comparably to the existing SAD measure. Specifically, the average PSNR results obtained by the proposed SAD-BM measure are only about 0.5 dB and 0.4 dB inferior to those obtained by the existing SAD measure when applying the FS and TSS algorithms, respectively. The computational gains compared with the conventional SAD measure in conjunction with the FS and TSS algorithms are 36 and 21, respectively.

References

1. J. R. Jain, and A. K. Jain, “Displacement measurement and its applications in interframe image coding," *IEEE Trans. on Communications*, vol. 29, no. 12, pp. 1799–1808, Dec 1981
2. R. Liu, B. Zeng, and M. L. Liou, “A new three-step search algorithm for block motion estimation", *IEEE Trans. Circ. Sys. Video Tech.*, vol. 4, no. 4, pp. 438-441, Aug 1994.
3. C. H. Cheung and L. M. Po, “A novel cross-diamond search algorithm for fast block motion estimation," *IEEE Trans. Circ. Sys. Video Tech.* , vol. 12, no. 12, pp. 1168–1177, Dec 2002
4. T. Koga, K. Linuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion compensated interframe coding for video conferencing", *Proc. of the Natl. Telecomm. Conf.*, pp. G.5.3.1–5.3.5, 1981
5. J. S. Kim and R. H. Park, “Feature-based block matching algorithm using integral projections", *IEEE Journal on Selected Areas in Comm.*, vol. 10, no. 5, pp. 968–971, June 1992
6. P. Viola and M. J. Jones, “Robust real-time object detection,” *Tech. Rep. CRL 2001/01*, Cambridge Research Laboratory, Feb 2001
7. B. Shen, I. K. Sethi and B. Vasudev, “Adaptive motion-vector resampling for compressed video downscaling”, *IEEE Trans. Circ. Sys. Video Tech.*, vol. 9, pp. 929–936, Sep 1999
8. MPEG Software Simulation Group (MSSG), “Test Model 5 (TM5)” [Online] Available: http://www.mpeg.org/MPEG/MSSG/

Impact of Similarity Threshold on Arbitrary Shaped Pattern Selection Very Low Bit-Rate Video Coding Algorithm

Manoranjan Paul and Manzur Murshed

Gippsland School of Computing and Information Technology,
Monash University, Churchill Vic 3842, Australia
{manoranjan.paul,manzur.murshed}@infotech.monash.edu.au

Abstract. Very low bit-rate video coding using arbitrary shaped patterns to represent moving regions in macroblocks has very good potential for improved coding efficiency. For any pattern based coding similarity threshold is used as a matching criterion between a moving region and a pattern. This metric together with quantization can control the coding efficiency curve. Unlike the quantization step size, the benefit of this metric is that it does not need to be transmitted any information in the decoder. Finer changes of coding efficiency curve can be possible by changing the similarity threshold instead of changing the quantization level, as a result a number of bits will be reduced. In this paper, we investigate the coding efficiency curves of different similarity thresholds.

1 Introduction

Reducing the transmission bit-rate while concomitantly retaining image quality continues to be a challenge for efficient video compression standards, such as H.263 [5], MPEG-2 [3].These standards are however inefficient while coding at very low bit-rate (VLBR) ($\leq$ 64 Kbps) due to inability to encode moving objects within a 16×16 pixel macroblock (MB) during motion estimation (ME), resulting in all 256 residual error values being transmitted for motion compensation (MC) regardless of whether there are moving objects. H.264/AVC standard [6] extended this block based motion compensated coding idea by introducing variable-block size (from 16×16 to 4×4) to approximate the shape of the moving objects within the MB more accurately. It requires a separate motion vector for each partition and the choice of partition size and the number of partition types has a significant impact on coding efficiency. It can be easily observed that the possibility of choosing smaller partition sizes diminishes as the target bit rate is lowered. Consequently, the coding efficiency improvement due to MB partitioning can no longer be realized for a VLBR target as larger partition sizes have to be chosen in most of the cases to keep the bit-rate in check at the expense of inferior shape approximation.

To address this problem, Fukuhara et al. [1] first proposed pattern based coding using four MB-partitioning patterns of 128-pixels each. By treating identically each MB, irrespective of its motion content, also resulted in a higher

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 184–191, 2004.

bit-rate being incurred for those MBs which contained only static background or had moving object(s), but with little static background. In such cases, the motion vectors for both partitions were almost the same and so only one could be represented.

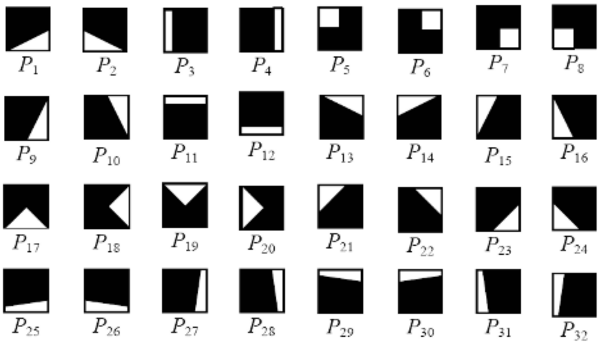


Fig. 1. The pattern codebook of 32 regular shaped, 64-pixel patterns, defined in 16 × 16 blocks, where the white region represents 1 (motion) and black region represents 0 (no motion).

The MPEG-4 [4] video standard first introduced the concept of content-based coding, by dividing video frames into separate segments comprising a background and one or more moving objects. To address the limitations of [1], Wong et *al.* [14] exploited the idea of partitioning the MBs via a simplified segmentation process that again avoided handling the exact shape of moving objects, so that popular MB-based motion estimation techniques could be applied. Wong et *al.* classified each MB into three distinct categories: 1) Static MB (SMB): MBs that contain little or no motion; 2) Active MB (AMB): MBs which contain moving object(s) with little static background; and 3) Active-Region MB (RMB): MBs that contain both static background and part(s) of moving object(s). SMBs and AMBs are treated in exactly the same way as in H.26X. For RMB coding, Wong assumed that the moving parts of an object may be represented by one of the eight predefined patterns $P_1 - P_8$ in Figure 1. An MB is classified as RMB if by using some similarity measure, the part of a moving object of an MB is well covered by a particular pattern. The RMB can then be coded using the 64 pixels of that pattern with the remaining 192 pixels being skipped as static background. Successful pattern matching can theoretically therefore have a maximum compression ratio of 4:1 for any MB. The actual achievable compression ratio will be lower due to the computing overheads for handling an additional MB type, the pattern identification numbering and pattern matching errors.

Other pattern matching algorithms have been reported [8]–[12]. Figure 1 shows the complete 32-pattern codebook. The performance of the RTPS algorithm [10] has been shown to be superior to all existing pattern matching algo-

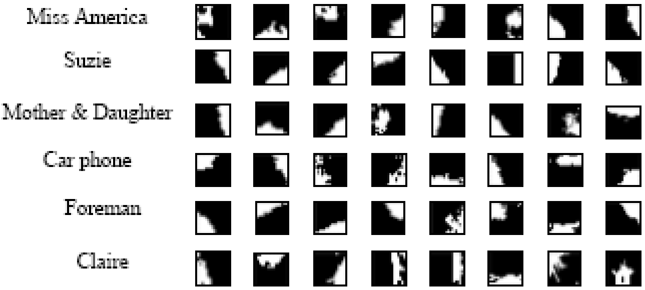


Fig. 2. Patterns extracted from video sequences by ASPS algorithm.

rithms. RTPS(4) for example, improved the peak signal to noise ratio (PSNR) value by up to 0.81dB compared with the Fixed-8 [14] algorithm and up to 1.52dB in comparison with H.263.

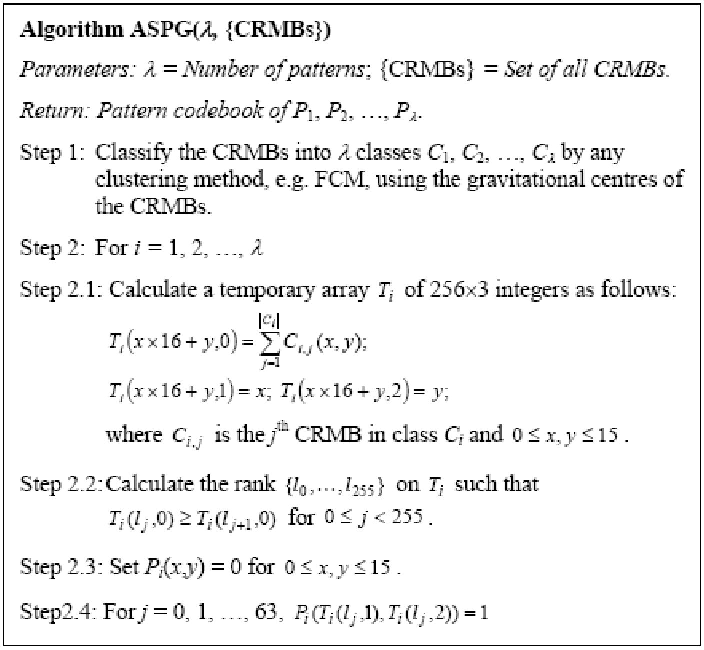
Paul et al. proposed an efficient Pattern Excluded Similarity Metric [12] and a content based Arbitrary Shaped Pattern Selection (ASPS) algorithm [11] which firstly extracted patterns from the actual video content without assuming any pre-defined shape and then used these extracted patterns to represent the RMB using a similarity measure as in all other pattern matching algorithms. Figure 2 shows the patterns generated by ASPS algorithm from some standard video sequences.

The ASPS algorithm like any other pattern based coding algorithm uses a similarity threshold to match a moving region with a pattern. ASPS algorithm with larger similarity threshold can capture more RMBs and as a consequence the bit-rate will be lower and the image quality will also be lower. On the other hand ASPS algorithm with smaller similarity threshold will capture less number of RMBs and as a result the bit-rate will be higher with image quality. In this paper we investigate the performance of ASPS algorithm for various similarity thresholds.

This paper is organized as follows. The video coding strategy using the ASPS algorithm is described in Section 2, while some simulation results are analysed in Section 3. Importance of similarity threshold is discussed in Section 4. Some future works and conclusions are provided in Section 5.

2 Pattern Based VLBR Coding

Prior to video coding, a pattern codebook (PC) has to be constructed. The ASPS algorithm performs this in two phases. In first phase, the PC is formulated on the basis of the actual video content, while in the second phase, the coding is undertaken using this content-dependent PC.



```
Algorithm ASPG($\lambda$, {CRMBs})
Parameters: $\lambda$ = Number of patterns; {CRMBs} = Set of all CRMBs.
Return: Pattern codebook of $P_1, P_2, \ldots, P_\lambda$.
Step 1: Classify the CRMBs into $\lambda$ classes $C_1, C_2, \ldots, C_\lambda$ by any
        clustering method, e.g. FCM, using the gravitational centres of
        the CRMBs.
Step 2: For $i = 1, 2, \ldots, \lambda$
Step 2.1: Calculate a temporary array $T_i$ of 256×3 integers as follows:
          $T_i(x \times 16 + y, 0) = \sum_{j=1}^{|C_i|} C_{i,j}(x,y)$;
          $T_i(x \times 16 + y, 1) = x$; $T_i(x \times 16 + y, 2) = y$;
          where $C_{i,j}$ is the $j^{th}$ CRMB in class $C_i$ and $0 \le x, y \le 15$.
Step 2.2: Calculate the rank $\{l_0, \ldots, l_{255}\}$ on $T_i$ such that
          $T_i(l_j, 0) \ge T_i(l_{j+1}, 0)$ for $0 \le j < 255$.
Step 2.3: Set $P_i(x,y) = 0$ for $0 \le x, y \le 15$.
Step2.4: For $j = 0, 1, \ldots, 63$, $P_i(T_i(l_j,1), T_i(l_j,2)) = 1$
```

Fig. 3. The ASPG algorithm.

2.1 PC Generation

Let $C_k(x,y)$ and $R_k(x,y)$ denote the kth MB of the current and reference frames, each of size $Wpixels \times Hlines$, respectively of a video sequence, where $0 \le x, y \le 15$ and $0 \le k < W/16 \times H/16$. The moving region $M_k(x,y)$of the kth MB in the current frame is obtained as follows:

$$M_k(x,y) = T(\mid C_k(x,y) \bullet B - R_k(x,y) \bullet B \mid) \tag{1}$$

where B is a 3×3 unit matrix for the morphological closing operation $\bullet$ [2] [7], which is applied to reduce noise, and the thresholding function T(v) = 1 if $v > 2$ and 0 otherwise.

If $8 \le \sum M_k < T_S + 64$ where T_S is a similarity threshold, then the kth MB is defined as a candidate RMB (CRMB). The Arbitrary Shaped Pattern Generation (ASPG) algorithm detailed in Figure 3 then generates the PC of $P_1, \ldots, P_\lambda$ using all CRMBs and user-defined pattern size λ. Any clustering method, such as Fuzzy C-Means (FCM) can be used in the ASPG algorithm. The clustering method classifies all CRMBs into classes using the gravitational centre (GC), which is defined as follows: Let G(A) be the GC of a 16×16 binary matrix A, such that

$$G(A) = \frac{\sum_{x=0}^{15}\sum_{y=0}^{15} xA(x,y)}{\sum_{x=0}^{15}\sum_{y=0}^{15} A(x,y)}, \frac{\sum_{x=0}^{15}\sum_{y=0}^{15} yA(x,y)}{\sum_{x=0}^{15}\sum_{y=0}^{15} A(x,y)} \tag{2}$$

By using FCM, those CRMBs with less inter GC distance are placed in the same class. The ASPS algorithm then adds all the corresponding '1's of those CRMBs in the same class to provide the most populated moving region. To create the 64-most populated moving regions as a pattern, only the first 64-pixel positions are assigned '1' with all the rest assigned '0'.

```
Algorithm PBC(PC)

Parameters: PC is the given pattern codebook.Return: Coded bitstream.

For each frame to be coded with motion compensation
    For each $k$-th MB in the current frame
        If $|M_k|_1 < 8$ then classify the block as SMB and skip from coding.
        Else if $8 \leq |M_k|_1 < T_S + 64$ and (4) is satisfied then classify the
            block as RMB and code the index of pattern $P_i$ and the moving
            region covered by this pattern using ME and MC while static
            region is skipped.
        Else classify the block as an AMB and code it using full ME and
            MC as is done in H.264.
```

Fig. 4. The general Pattern based video coding (PBC) algorithm.

2.2 Actual Coding

Let $|Q|_\ell$ be the total number of ℓ 's in the matrix Q. Similarity of a pattern $P_n \in PC$ with the moving region in the kth MB can be defined efficiently [12] as

$$S_{k,n} = |M_k|_1 - |M_k \wedge P_n|_1 \tag{3}$$

Clearly, higher the similarity lower will be the value of $S_{k,n}$. The CRMB is classified as an RMB and its moving region is represented by a pattern P_i such that

$$P_i = \arg \min_{\forall P_n \in PC} (S_{k,n} | S_{k,n} < T_S) \tag{4}$$

where T_S is the predefined similarity threshold; otherwise the CRMB is classified as an AMB.

For a given PC, an image sequence is coded using the general pattern based coding (PBC) algorithm in Figure 4. To avoid more than one 8×8 block of DCT calculations for 64 residual error values per RMB, these values are rearranged into an 8×8 block. It avoids unnecessary DCT block transmission. A similar inverse procedure is performed during the decoding.

3 Simulation Results

To compare the performance of both the ASPS algorithm with different similarity thresholds and H.264 standard we tested a large number of standard and non-standard video sequences of QCIF digital video formats [13]. For the purposes of this paper, experimental results are presented using the first 100 frames of four standard video test sequences. Full-search, half-Pel, and variable block-size ME and MC were employed to obtain the encoding results using the ASPS approach and H.264 standard. The ASPS algorithm used $\lambda = 8$.

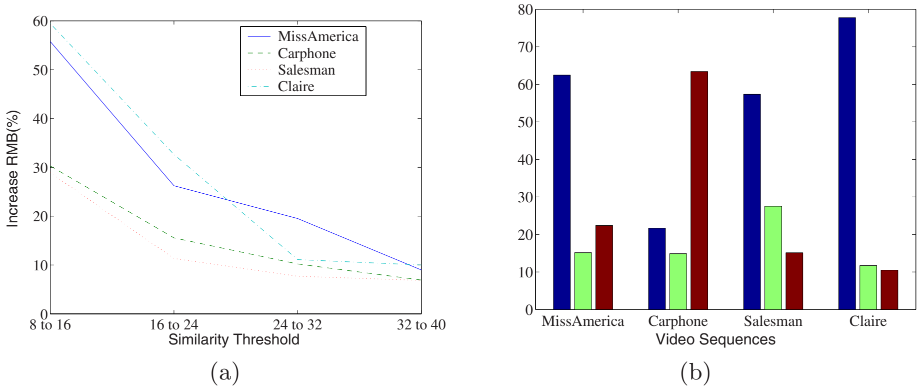


Fig. 5. (a) Percentage of increased RMBs by ASPS algorithm while similarity thresholds are changes from 8 to 16, 16 to 24, 24 to 32, and 32 to 40; (b) Percentage of SMBs, RMBs, and AMBs by ASPS algorithm where similarity threshold is 16.

Figure 5(a) shows the percentage of increased RMBS by ASP algorithm when similarity threshold changes from 8 to 16, 16 to 24, 24 to 32, and 32 to 40. We observed the diminishing trends of increasing the RMBs when the similarity threshold is already large. The original percentages of different MBs generated by ASPS algorithm are shown in Figure 5 (b) where similarity threshold is 16. Note that a rough idea about the motion involvement in a particular video sequences can be concluded by observing the relative number of MB types. For example, the motion involvement of carphone is much greater than salesman sequence as the AMB and RMB of carphone are larger than that of Salesman.

The coding performance of the ASPS algorithm like any other pattern based video coding algorithm depends on the value of the similarity threshold. The ASPS algorithm with larger similarity threshold can capture more RMBs and as a consequence the bit-rate will be lower and the image quality will also be lower. On the other hand ASPS algorithm with smaller similarity threshold will capture less number of RMBs and as a result the bit-rate will be higher with image quality. Figure 6 shows the coding efficiency curves by ASPS algorithm with various similarity threshold denoted by parameter as well as H.264 standard.

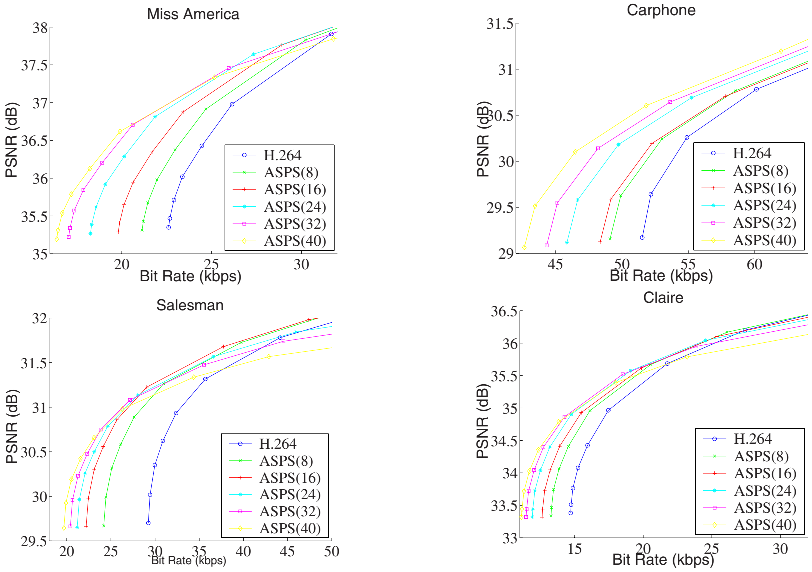


Fig. 6. Coding performance comparisons for four standard test video sequences by ASPS algorithm with various similarity thresholds denoted by parameter and H.264 standard.

4 Importance of Similarity Threshold

The similarity threshold has a greater role in controlling the bit-rate over a limited bandwidth channel. Normally a video coding algorithm control the bit rate by increasing or decreasing the Quantization level and amount of quantization changes are coded together with video data. In Figure 7 we observed that different similarity thresholds provide different bit rates. But no bits are needed to send in the decoder end about the changes of the similarity threshold. Thus, an adaptive ASPS algorithm can control the bit rate by changing the similarity thresholds instead of changes the quantization level and as a result a large number of bits will be saved. However, the different quantization level is needed where a large change of bit-rate is required.

5 Future Works and Conclusions

Video coding using arbitrary shaped patterns to represent the moving region in macroblocks performed better than the H.264 standard especially for very low bit rate video coding because the former represents an MB by a smaller size moving region covered by the best available pattern that approximates the shape of the region more closely and hence, requiring no extra motion vector, which is not the case with the latter. For any pattern based coding including ASPS algorithm, similarity threshold is used as a matching criterion between a moving region and a pattern. This metric together with quantization level can control the coding efficiency curve. Unlike the quantization step size, the benefit of this metric is that it does not need to be transmitted any information in the decoder end. Finer changes of coding efficiency curve can be possible by changing the similarity threshold instead of changing the quantization level as a result a number of bits will be reduced. But the existing ASPS algorithm cannot select the suitable similarity threshold for the channel requirements. We are investigating to design an adaptive ASPS algorithm which can utilize the various similarity thresholds to control the bit-rate instead of quantization level in finer adjustments.

References

1. Fukuhara, T., K. Asai, and T. Murakami: Very low bit-rate video coding with block partitioning and adaptive selection of two time-differential frame memories. IEEE Trans. Circuits Syst. Video Technol., Vol. 7, pp. 212–220, 199
2. Gonzalez, R.C. and R. E. Woods: Digital Image Processing. Addison-Wesley, 1992
3. ISO/IEC 13818, MPEG-2 International Standard, 1995
4. ISO/IEC N4030, MPEG-4 International Standard, 2001
5. ITU-T Recommendation H.263: Video coding for low bit-rate communication. Version 2, 1998
6. ITU-T Rec. H.264/ISO/IEC 14496-10 AVC. Joint Video Team (JVT) of ISO MPEG and ITU-T VCEG, JVT-G050, 2003
7. Maragos, P.: Tutorial on advances in morphological image processing and analysis. Opt. Eng., Vol. 26 no. 7, pp. 623–632, 1987
8. Paul, M., M. Murshed, and L. Dooley: A Low Bit-Rate Video-Coding Algorithm Based Upon Variable Pattern Selection. Proc. of 6th Int. Conf. on Signal Processing (ICSP-02), Beijing, Vol-2, pp. 933–936, 2002
9. Paul, M., M. Murshed, and L. Dooley: A new real-time pattern selection algorithm for very low bit-rate video coding focusing on moving regions. Proc. of IEEE Int. Conference of Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, Vol-3, pp. 397–400, 2003
10. Paul, M., M. Murshed, and L. Dooley: A Real-Time Pattern Selection Algorithm for Very Low Bit-Rate Video Coding Using Relevance and Similarity Metrics. To appear in IEEE trans. on circuits and systems on video technology.
11. Paul, M., M. Murshed, and L. Dooley: An Arbitrary Shaped Pattern Selection Algorithm for VLBR Video Coding Focusing on Moving Regions. Proc. of 4th IEEE Pacific-Rim Int. Con. on Multimedia (PCM-03), Vol. 1, pp. 100–104, 2003
12. Paul, M., M. Murshed, and L. Dooley: A New Efficient Similarity Metric and Generic Computation Strategy for Pattern-based VLBR Video Coding. Proc. of the IEEE Int. Con. of Acoustics, Speech, and Signal Proc. (ICASSP-04), 2004
13. Shi, Y.Q. and H. Sun: Image and Video Compression for Multimedia Engineering Fundamentals, Algorithms, and Standards, CRC Press, 1999
14. Wong, K.-W., K.-M. Lam, and W.-C. Siu: An Efficient Low Bit-Rate Video-Coding Algorithm Focusing on Moving Regions. IEEE Trans. on Circuits and Systems for Video Technology, Vol. 11, no. 10, pp. 1128–1134, 2001

A Study on the Quantization Scheme in H.264/AVC and Its Application to Rate Control*

Siwei Ma[1], Wen Gao[1], Debin Zhao[1], and Yan Lu[2]

[1] Institute of Computing Technology,
Chinese Academy of Science, Beijing, China
{swma,wgao,dbzhao}@jdl.ac.cn
[2] Microsoft Research Asia, Beijing, China
t-yanlu@microsoft.com

Abstract. Compared with previous video coding standards, H.264/AVC employs a division-free quantization scheme. The relation between quantization parameter and quantization step changes from general linear to exponential as well. In this paper, we first analyze the relation between rate and quantization step-size and derive a new rate-distortion (R-D) model. An efficient rate control scheme is then developed based on the new R-D model. The proposed rate control scheme is implemented into the H.264/AVC reference software, with which the better coding performance can be achieved. Experimental results show that the PSNR of the proposed rate control scheme is averagely 0.23dB over the current rate control scheme in H.264/AVC test model and 0.41dB over the coding scheme with fixed quantization parameter, and meanwhile, the complexity of the proposed rate control scheme is much lower than the original one.

1 Introduction

Rate control plays an important part in any standard-compliant video codec. Without rate control, any video coding standard would be practically useless. As a conse-quence, a proper rate control scheme was usually recommended for a standard during the development, e.g. TM5 [1] for MPEG-2, TMN8 [2] for H.263 and VM8 [3] for MPEG-4, etc. H.264/AVC is the newest international video coding standard, and some work about rate control has been done for H.264/AVC too. In [5], a rate control scheme based on VM8 has been proposed to and adopted by H.264/AVC test model. In our previous work [6], rate distortion optimization and hypothetical reference decoder (HRD) have been jointly considered in rate control implementation process, part of which has also been adopted by H.264/AVC test model.

Generally, the rate control can be implemented in two steps. The first step is to allocate appropriate bits for each picture. The bit allocation process is

* This paper has been partly supported by NSFC(60333020) and National Hi-Tech Research Program(2002AA119010).

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 192–199, 2004.

constrained by a HRD model defined in the standard specification. The second step is to adjust quantization parameter (QP) for each coding unit (e.g. the macroblock) so as to fulfill the target bitrate constraint. In other words, the key point is to find the relation between the rate and QP. Since the source distortion is closely related with the quantization errors as well as the QP, the relation between rate and QP is usually developed based on a rate-distortion (R-D) model. For example, in TM5 a simple linear rate-distortion model is introduced. In TMN8 and VM8, the more accurate R-D models are used, which can reduce rate control error and provide better performance but have relatively higher computational complexity. In [4,7], the relation between rate and QP is indirectly represented with the relation between rate and , where is the percent of zero coefficients after quantization. In [8] and [9], a modified linear R-D model with an offset indication overhead bits is used for rate control on H.26x.

In conclusion, these rate control schemes are mostly associated with the true relation between rate and quantization parameters in the video codec. Since many coding tools in H.264/AVC, in particular the quantization scheme, differ from the previous video coding standards, it is desirable to derive the new relation between the rate and distortion as well as QP for rate control on H.264/AVC. This paper is an extension and refinement of our previous researches. The quantization scheme in H.264/AVC is first fully studied to reveal the true relation between rate and quantization parameter and derive the new R-D model. Based on this new R-D model, a macroblock-layer rate control is proposed and implemented on the H.264/AVC reference software while with lower complexity compared with current rate control scheme [5] in H.264/AVC. Because the complicated MAD prediction and R-D model in [5] makes it difficult to be used in real time encoder.

The rest of the paper is organized as follows. Section 2 presents the detailed analysis on the quantization scheme in H.264/AVC, and then describes the proposed R-D model. In Section 3, the strategy of the proposed rate control algorithm is described. The experimental results are presented in Section 4. And finally, Section 5 concludes this paper.

2 Quantization Scheme in H.264/AVC

Before we present the proposed rate control algorithm, we first make a study on the quantization scheme and the rate-distortion relation in H.264/AVC. Above all we should distinguish QP and quantization step-size (referred to as Q_{step} hereafter). QP denotes the quantization scale indirectly, whereas Q_{step} is the true value used in quantization. In the previous video coding standards, the relation between QP and Q_{step} is usually linear. For example, in H.263 quantization scheme, in terms of the quantization parameter QP, a coefficient COF is quantized to:

$$LEVEL = \left\{ \begin{array}{c} |COF|/(2 \times QP) \\ |COF - QP/2|/(2 \times QP) \end{array} \right. \tag{1}$$

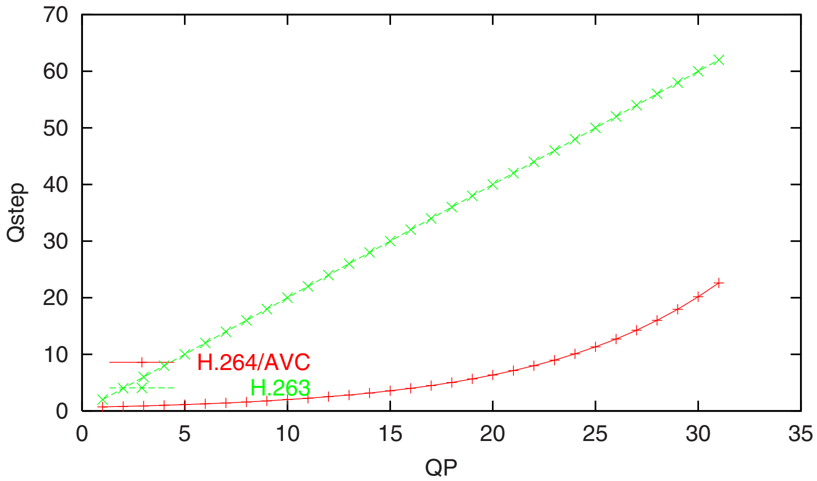


Fig. 1. The relation between QP and Qstep in H.264/AVC

where quantization step-size $Q_{step} = 2QP$. However, in H.264/AVC, the relation between QP and Q_{step} is that $Q_{step} = 2^{(QP/6)}$, as shown in Fig. 1. The underlying reason for this change is that an integer transform and division free quantization scheme is adopted by H.264/AVC. In H.264/AVC, the following integer transform is used to do transformation [10]:

$$Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \begin{bmatrix} X \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 1 & -1 & -2 \\ 1 & -1 & -1 & 2 \\ 1 & -2 & 1 & -1 \end{bmatrix} \tag{2}$$

Since the integer transform is not a unitary matrix, Y must be normalized as follows:

$$W = \begin{bmatrix} Y \\ \\ \end{bmatrix} \otimes \begin{bmatrix} E \\ \\ \end{bmatrix} = \begin{bmatrix} Y \\ \\ \end{bmatrix} \begin{bmatrix} a^2 & ab & a^2 & ab \\ ab & b^2 & ab & b^2 \\ 1 & -1 & -1 & 2 \\ 1 & -2 & 1 & -1 \end{bmatrix} \tag{3}$$

where a=1/2, b=$1/\sqrt{10}$. In H.264/AVC, in terms of quantization parameter QP, the transform normalization combining with the quantization is implemented for division free as:

$$LEVEL_{i,j} = round(W_{i,j}/Q_{step}) = round(Y_{i,j}S_{i,j}/2^{qbits}) \tag{4}$$

where $S_{i,j}/2^{qbits} = E_{i,j}/Q_{step}$, $Q_{step} = 2^{((QP-4)/6)}$ and $qbits = 15 + floor(QP/6)$.

To model the relation between the rate R and the quantization step-size Q_{step} as well as quantization parameter QP, we make the statistics as follows. Figure 2 shows the relations of $R-(1/QP)$ and $R-(1/Q_{step})$ for the News

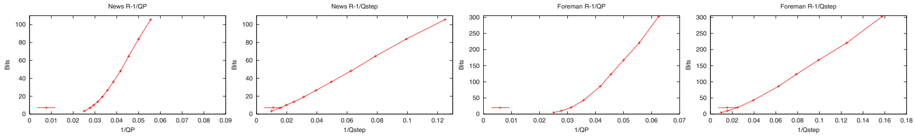


Fig. 2. The relation between R and 1/QP(1/Qstep) on News and Foreman.

sequence, respectively, wherein R denotes the bits for coefficients of luma and chroma components. Table 1 shows the statistics on the Foreman sequence. The correlation factor $\rho_{R-(1/QP)}$ for R and $1/QP$ is 0.991, and the correlation factor $\rho_{R-(1/Qstep)}$ for R and Q_{step} is 0.999. Though R and $1/QP$ is also highly linear-correlated, the sum of squared error (R'_i is the linear aproximation of R_i according to the linear model resolved by least square error multiplier) for the linear approximation of $R-(1/QP)$ is much larger than that of $R-(1/Q_{step})$. Test results on other sequences also show the similar statistical result. More experiments on some other sequences also prove the similar results.

Table 1. Experimental results on News

QP	18	20	22	24	26	28	30	32	34	36	40
R	105.7	83.88	64.65	48.11	36.07	26.53	19.17	13.69	10.05	6.944	3.324
$\rho_{R-1/QP}$	0.991										
$\rho_{R-1/Q_{step}}$	0.999										
$\sum(R-R')^2_{1/QP}$	236.003										
$\sum(R-R')^2_{1/Q_{step}}$	15.1208										

Therefore, we can draw a conclusion that the relation between R and $1/Q_{step}$ can be thought as linear in H.264/AVC. The $R-Q_{step}$ model is then derived as:

$$R_i^t = K^t SAD_i/Q_{stepi} + C^t, t = I, P, B \tag{5}$$

where R_i^t is the estimated number of bits of a macroblock, SAD_j is the sum of absolute difference of a motion compensated macroblock. The first item reflects the bits used to code transform coefficients. The second item is the bits used to code header information of a macroblock. Compared with the linear R-D model, e.g. *R-QP* model in TM5, the $R-Q_{step}$ model is more accurate for H.264/AVC due to the new quantization scheme.

3 Rate Control Algorithm

In the previous section, a linear $R-Q_{step}$ model has been proposed to reveal the relation of rate and quantization step-size in H.264/AVC. In this section, an efficient rate control scheme for H.264/AVC is presented. Concretely, the proposed rate control algorithm is performed as follows:

Step 1. Bit allocation.
In this step, a target bit is allocated to each picture in a group of picture (GOP) as TM5. For the first I/P/B picture, bit allocation is not performed, and a fixed QP_0^t is then used. After coding the picture, parameters X_0^t, K^t and C^t are initialized, respectively.

$$X_0^t = S_0^t {Q_{step}}_0^t = S_0^t 2^{((QP_0^t - 4)/6))}, C^t = B_{head} \; and$$
$$K^t = B_{coeff} {Q_{step}}_0^t / SAD_t = B_{coeff} 2^{((QP_0^t - 4)/6))} / SAD_t \tag{6}$$

S_0^t is the coded bits of the frame; SAD_t is the average SAD of all macroblocks in the frame; B_{head} is the average header bits for a macroblock, including motion and mode information; and B_{coeff} is the bits used to code luma and chroma coefficients. Set $j^t = 0$ for the first picture with type t, which is used to update K^t in the R-D model. Set $K_1 = KK_0 = K^t, C_1 = CC_0 = C^t$. N^t is the number of t type pictures in the current GOP.

Step 2. Initialization of the current macroblock.
In this step, we initialize some parameters. Let $i = 1$ for the first macroblock. Assume B_1 is the number of available bits for coding this frame, and L_1 is the number of remained not coded macroblocks in the current frame.

Step 3. Rate distortion optimization mode selection for the current macroblock.
If the current macroblock is the first one in a frame, QP is set to be the average quan-tization parameter QP_{prev} of previous frame; otherwise, if $B_i/L_i < C^t, QP = QP_{prev}$, else a new Q_{step} for the current macroblock is calculated as:

$$Q_{step} = K^t SAD_i - 1/(B_i/L_i - C^t) \tag{7}$$

Therefore, $QP = round(6log_2 Q_{step}) + 4$. QP is then clipped with:

$$QP = min(max(QP_{prev} - 3, QP), QP_{prev} + 3) \tag{8}$$

QP must be clipped to be in the range from 0 to 51. The new QP is used in the coding mode selection of the current macroblock.

Step 4. Counter updating.
In this step, the remaining bits and the number of not coded macroblocks of the frame are updated as follows:

$$B_{i+1} = B_i - R_i, and L_{i+1} = L_i - 1$$

Assume MB_CNT is the total number of macroblocks in the frame. If $i = MB\_CNT$, all macroblocks in the frame are coded; otherwise, let $i = i + 1$, and go to Step 2.

Step 5. R-D model parameter updating.
The R-D mode parameters K and C are updated similar to [2], but at frame level. First calculate:

$$K' = \frac{R_{C,n} Q_{step}}{SAD_n^t MB\_CNT}, C' = \frac{R_n - R_{C,n}}{MB\_CNT} \tag{9}$$

where $R_{C,n}$ is the number of bits spent for the luminance and chrominance of the nth picture with picture type t. If $(K' >= 0 and K' <= threshold)$, set $j^t = j^t + 1$ and:

$$KK_{j^t} = KK_{j^t-1}(j^t - 1)/j^t + K'/j^t, CC_n = CC_{n-1}/n + C'/n \tag{10}$$

K^t and C^t are updated as a weighted average of the initial estimates and the current average:

$$K^t = KK_{j^t} n/N^t + K_1(N^t - n)/N, C^t = CC_n n/N^t + C_1(N^t - n)/N^t \tag{11}$$

4 Experiments and Results

In order to evaluate the performance of the proposed algorithm, some experiments have been done on the typical test sequences. Three rate control schemes, i.e. the proposed rate control scheme, the modified TM5 rate control scheme with some parameters adjustment for H.264/AVC and the rate control scheme currently adopted in H.264/AVC test model are implemented on the same H.264/AVC reference software, respectively. The comparisons are then performed among these rate control schemes and fixed QP coding.

The experimental results are listed in Table 2. According to the table, we can see that the proposed algorithm can effectively control the bit-rate at different resolution, frame rate, and meanwhile it can also achieves better coding efficiency than the rate control schemes and fixed QP coding. The maximum improvement compared to the current rate control in H.264/AVC test model [5]

Table 2. Experimental results on test sequence

Sequence	Squence Type	Target Bit-rate (kbps)	Rate Control Scheme	Coded Bit-rate (kbps)	PSNRY (dB)	I mprovement over (dB)		
						[5]	TM5	Fixed QP
News	QCIF 10f/s, IPP	10.19	Proposed	10.19	28.78	0.48	0.98	0.78
			[5]	10.23	28.30			
			TM5	10.20	27.80			
			None(Fixed-QP)	10.19	28.00			
Foreman	QCIF 15f/s, IBBP	98.00	Proposed	97.76	36.35	0.06	0.50	0.25
			[5]	97.76	36.29			
			TM5	98.00	35.85			
			None(Fixed-QP)	98.00	36.10			
Container	QCIF 10f/s, IPP	12.84	Proposed	12.92	33.56	0.15	0.71	0.46
			[5]	12.99	33.41			
			TM5	13.15	32.85			
			None(Fixed-QP)	12.84	33.10			
Bus	QCIF 30f/s, IPP	93.84	Proposed	93.87	30.78	0.24	0.53	0.13
			[5]	93.88	30.54			
			TM5	94.36	30.23			
			None(Fixed-QP)	93.84	30.65			

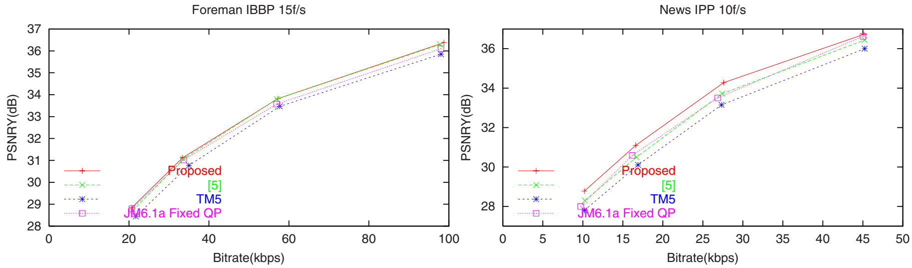


Fig. 3. PSNR curve of News at 10f/s and Foreman at 15fps, QCIF

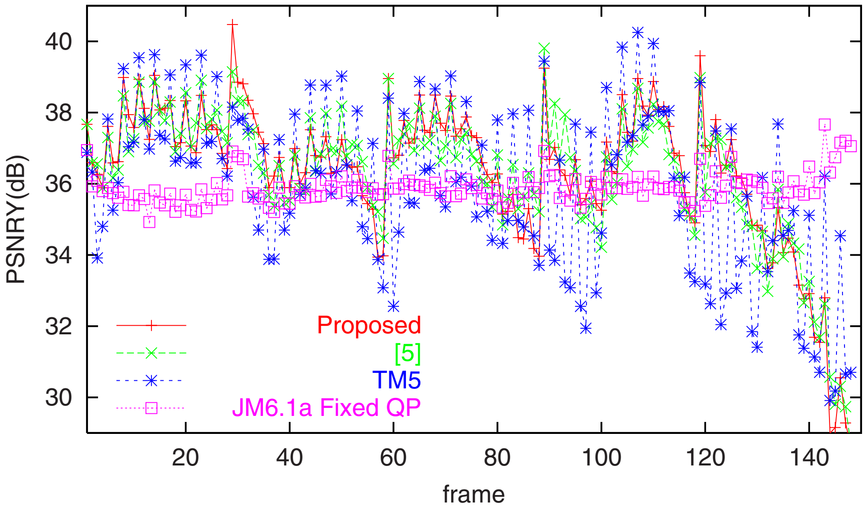


Fig. 4. PSNR per frame for Foreman 98.0kbps, 15f/s, QCIF

is 0.48dB and the average improvement reaches 0.16dB. Since the proposed algorithm only employs a simple linear model without performing the complicated MAD prediction, the complexity is also much lower than the current rate control in H.264/AVC test model. Compared with fixed *QP* coding, the proposed rate control can further improve coding efficiency with the average PSNR improvement reaching at 0.45dB. Compared with the modified TM5 implementation in H.264/AVC, the improvement is up to 0.98dB.

In Fig. 3, the rate-distortion curves of the proposed rate control, TM5, current rate control in H.264/AVC test model and fixed *QP* coding are shown, respectively. From the curves we can see that the proposed rate control can achieve better performance than any other schemes, which demonstrates that the R-D model in the proposed rate control scheme is more accurate for H.264/AVC video coding. Since the bits of coding the motion vectors play a part in the overall rate, the motion in the sequence may also influence the rate control scheme.

For the News sequence with small motion, the proposed rate control can achieve much better performance than the other schemes. The performance is almost the optimum. For the Foreman sequence with high motion, the performance of the proposed scheme is also a little better than the other schemes, which demonstrates that the proposed R-D model is still more accurate than the others even for the sequence with high motion.

Figure 4 shows the PSNR per frame in terms of the Foreman sequence. The figure further indicates that the proposed rate control scheme shows better performance and has improved the coding efficiency of the original H.264/AVC test model.

5 Conclusion

This paper has presented a detailed study on the quantization scheme in H.264/AVC, from which a more accurate rate distortion model has been derived. Based on the proposed rate distortion model, an efficient rate control scheme has been proposed for H.264/AVC video coding. The experimental results have shown that the proposed scheme can further improve the coding efficiency of the original H.264/AVC test model. In addition, the proposed rate control scheme can also outperform the current rate control scheme in H.264/AVC test model, and meanwhile its computational complexity is also lower.

References

1. ISO-IEC/JTC1/SC29/WG11, "Test model 5," JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio, MPEG 1994
2. J. Corbera and S. Lei, "Rate control for low-delay video communications," ITU Study Group 16, Video Coding Experts Group Documents Q15-A-20, Portland, June 1997
3. T. Chiang, Y. Zhang, "A new rate control scheme using quadratic rate distortion model," IEEE Trans. Circuits Syst. Video Technol. Vol. 7, pp. 287–311, Apr. 1997
4. S. Milani, L. Celetto, G. A. Mian, "A rate control algorithm for the H.264 encoder"
5. ISO IEC TC JTC 1/SC29 N5821, Draft ISO/IEC 14496-5:2002/PDAM6: 2003
6. S. Ma, W. Gao, F. Wu, Y. Lu, "Rate control for AVC video coding scheme with HRD considerations," IEEE International Conference on Image Processing, ICIP 2003, Barcelona, Spain, 14–17 Sept. 2003
7. Z. He, S. K. Mitra, "A unified rate-distortion analysis frame work for transform coding," IEEE Trans. on Circuits Syst. Video Technology. Vol. 11. No. 12, Dec. 2001
8. J. Wang, Z. Chen, Y. He, Y. Chen, "A MAD-based rate control strategy," Document JVT-D070, Klagenfurt, Austria, July 2002
9. C. Wong, O. C. Au, B. Meng, H. Lam, "Novel H.26X optimal rate control for low-delay communicaitons," ICICS-PCM 2003, 15–18 Dec. 2003, Sigapore
10. H.264/MPEG-4 Part 10: Transform & Quantization, http://www.vcodex.com

An Efficient VLSI Implementation for MC Interpolation of AVS Standard

Lei Deng[1], Wen Gao[2], Ming-Zeng Hu[1], and Zhen-Zhou Ji[1]

[1] Department of Computer Science and Engineering
Harbin Institute of Technology, Harbin 150001, China
lden@jdl.ac.cn

[2] Institute of Computing Technology, Chinese Academy of Science,
Beijing, 100080, China
wgao@jdl.ac.cn

Abstract. Advance Video Coding standard (AVS) [1] is the standard for compression and decompression in digital audio and video multimedia. The AVS Working Group was approved by the Science and Technology Department of Ministry of Information Industry of china on June 2002. AVS has employed a 4-tap interpolation FIR filter in its motion compensation (MC) part for high coding efficiency. But it is accompanied by increasing the complexity in calculation and memory access. And this problem makes MC one of the bottlenecks in the AVS system's VLSI implementation, especially for SDTV or HDTV which aggravate the problem heavily. Unfortunately, most FIR filter [3-5] have too low of input bandwidth to deal with it. In this paper, an efficient architecture for MC interpolation is described, and experimental results show that this architecture satisfies AVS decoder applications such as SDTV or HDTV.

1 Introduction

Compared with previous video coding standards like MPEG-2 and H.263, AVS obtains higher coding efficiency with advanced features and functionality like different block sizes, lagrangian coder control, multiple reference frames, fractional pixel precision, etc, at an increased implementation cost. The partition of the AVS's inter-coded block can be one of seven types identified by lumina block sizes: 16×16, 16×8, 8×16, 8×8. MC interpolation of the AVS employs a 4-tap FIR filter that needs the samples not only inside the current block but also outside it. This surely requires more memory bandwidth. The amount of memory access for MC interpolation is about 50% in the AVS decoder. And the time consumed by interpolation processing is about 25% in the AVS decoder major subsystems. So MC becomes one of the most data intensive parts of the AVS decoder, and a bottleneck of implementation. The memory access and the high transfer rate are the main troubles in the VLSI design of the MC interpolation.

According to the limit of AVS for SDTV or HDTV application level, the maximum macro-block (MB) rate is 245760 MB/s (30 frame/s and 8192 MB

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 200–206, 2004.

in a picture). If using 150MHz clock, the time budget is only 610 clock cycles assigned for the processing of one MB which has four luma blocks and two chroma blocks. And only slightly more than 100 cycles for a single block averagely.

Realization of FIR filters can vary widely from one that uses dedicated hardware multipliers and adders [3,4] to one that uses code executed by a general purpose processor. A combination of hardware and software that allows sharing of hardware units like adders and multipliers can also be used [5,7]. But these FIR filters don't fulfil the requirement of the interpolation in this paper. Extra clock cycles are required by mentioned FIR filters above between the processing for adjacent row or column of MB to displace the MB data inside their architectures. Extra clock cycles lead to the loss of processing time significantly, and cut down the filter efficiency. Moreover, these FIR filters have so insufficient data input bandwidth that they only allow to input one pixel data per cycle. For a 8×8 interpolated outcome the algorithm or the filters require to deal with 169 input pixel data [6], that means at least 169 cycles should be spent here. Thinking about the time budget of a single block—100 cycles, these FIR filters are so time-consuming that they can not meet the requirement. After analyzing their architecture thoroughly, we found that most of these FIR filters only have one-dimension data transfer way. But MC interpolation needs two-dimension MB data for calculation. So these FIR filters have low behavior on this interpolation. The data transfer scheme adopted by The AB2 type architecture [2,5] for motion estimation in video encoder is very similar to the requirement of interpolation in this paper. Compared with [2], Architecture of [5] reduced the hardware demands, and was implemented more economically.

The main goal presented in this paper is to arrange the MB data transfer properly and exploit the large bandwidth and high parallel architecture for MC interpolation of AVS decoder applications with larger frame sizes such as SDTV and HDTV. In section 2 of this paper, mc interpolation algorithm is described. Section 3 describes details of the implementation for the architecture of mc 4-tap interpolation. Experimental results are given in Section 4. The paper closes with a conclusion in Section 5.

2 MC 4-Tap Interpolation Algorithm

The aim of interpolation is to get the fractional samples from the integer samples according to the motion vector. In the AVS coding standard, the motion vector can point to the quarter-pel-accuracy location by the last two bits. Fig. 1 shows the positions of integer samples (Squares with uppercase letters) and fractional samples (Squares with lowercase letters) for MC interpolation inside the given two-dimensional sample array. In Fig. 1 Squares marked b, h, j, m and s are the same as those labeled with double lowercase letters such as aa which are positions at half-pel-accuracy location. And those others left and labeled with single lowercase letter are quarter-pel-accuracy locations. According to different locations pointed by motion vectors, fractional samples have different interpolation processing. Here, this processing is classified by five cases described below.

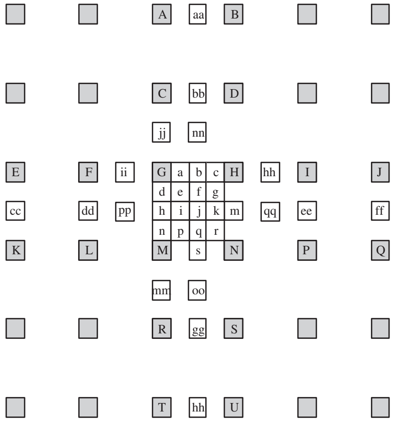


Fig. 1. The positions of integer samples and fractional samples

Case1: this is a special case, if the motion vector point to the integer sample of G, no processing needs to do here. And G is directly output.

Case2: this case includes half samples such as b, h, s and m, and so on. A 4-tap interpolation filter (-1, 5, 5, -1) needs as: $z' = -x_{-1} + 5x_0 + 5x_1 - x_2$, and $z = clip1((z' + 4) \gg 3)$. Where z' represents intermedia variables of b', h', s' or m', and z represents b, h, s or m. The subscripts of x index horizontal or vertical neighbouring integer-pixel locations. *clip1* stands for clipping between[0, 255]. They shall be available after one 4-tap filter execution time.

Case3: j, nn, oo, pp and qq belongs to this case. An example of this case, j shall be obtained as: $j' = -dd' + 5h' + 5m' - ee'$, and $j = clip1((j' + 32) \gg 6)$. Where, variables of dd', h', $m'$and ee' are available in the same manner of the intermedia variable z' in case2. Because both h' and j' are derived by 4-tap filter, j shall be obtained after two serial 4-tap filter execution times. So if j is the final result of the interpolation, the latency of hardware shall be twice as much as b.

Case4: this case includes e, g, p, and r, which use the half sample j'. And so they shall be derived after j is done (table1).

Case5: this case includes a, c, d, n, f, i, k, q, which at least use one filtered half sample such as b' in case2 or j' in case3. And a 4-tap interpolation filter (1, 7, 7, 1) needs as: $z' = -x_{-1} + 7x_0 + 7x_1 - x_2$, and $z = clip1((z' + 64) \gg 7)$(see table1). In table1, ii', hh', jj' and mm' are available in the same manner of the intermedia variable z' in case2, and all integer pixels need to be amplified eight times. Since the process for j needs two 4-tap filter execution times, three execution times may be used at latest in this case.

From the description above, it is clear to see the relations among the five cases. Firstly, fraction samples in case3, case5 depend on those in case2, and case4 and case5 depend on case3. Secondly, fraction samples in case4 and case3 have near the same interpolating complexity as twice much as those in case2, and case5 is the most complex case. These relations are very useful to optimize the design of the interpolation architecture.

Table 1.

case5	function	case4	funtion
a'	$ii' + 7 \times 8G + 7 \times b' + 8H$	e	$(64G + j' + 64) \gg 7$
c'	$8G + 7 \times b' + 7 \times 8H + hh'$		
d'	$jj' + 7 \times 8G + 7 \times h' + 8M$	g	$(64H + j' + 64) \gg 7$
n'	$8G + 7 \times h' + 7 \times 8M + mm'$		
f'	$nn' + 7 \times b' + 7 \times j' + s'$	p	$(64M + j' + 64) \gg 7$
i'	$pp' + 7 \times h' + 7 \times j' + m'$		
k'	$h' + 7 \times j' + 7 \times m' + qq'$	r	$(64N + j' + 64) \gg 7$
q'	$b' + 7 \times j' + 7 \times s' + oo'$		

3 The Architecture of 4-Tap Interpolation for MC

The data transfer scheme of the AB2 type architecture [5][2] is very suitable for the two-dimension data transfer process such as the MC interpolation. Base on it, the new architecture is devised and showed as Fig. 2. In the architecture the N and M is decided by the partition of the MB. Two parts are included in this architecture. One is the part of the pixel data transfer, and the other is the part of ALU.

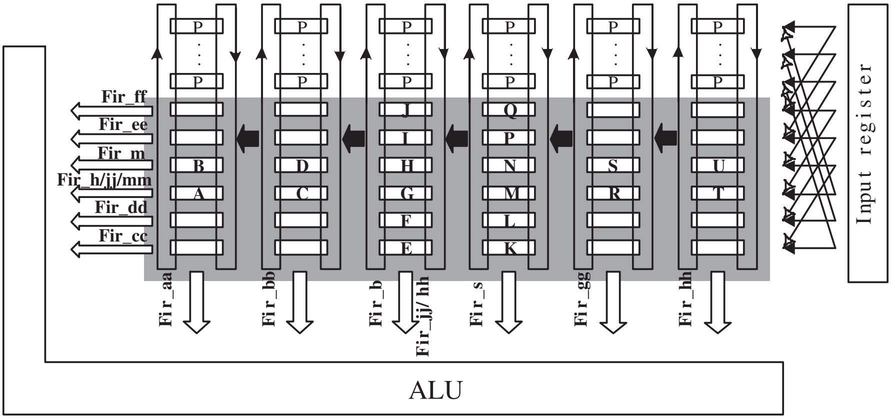


Fig. 2. The interpolation architecture of N X M block

3.1 Pixel Data Transfer Scheme

The scheme has a register array which has $N + 5$ row and 6 column registers used to reserve the data for the current or later processing. There are two type registers in the array, one is the active pixel register (A_{ij} register) which serves

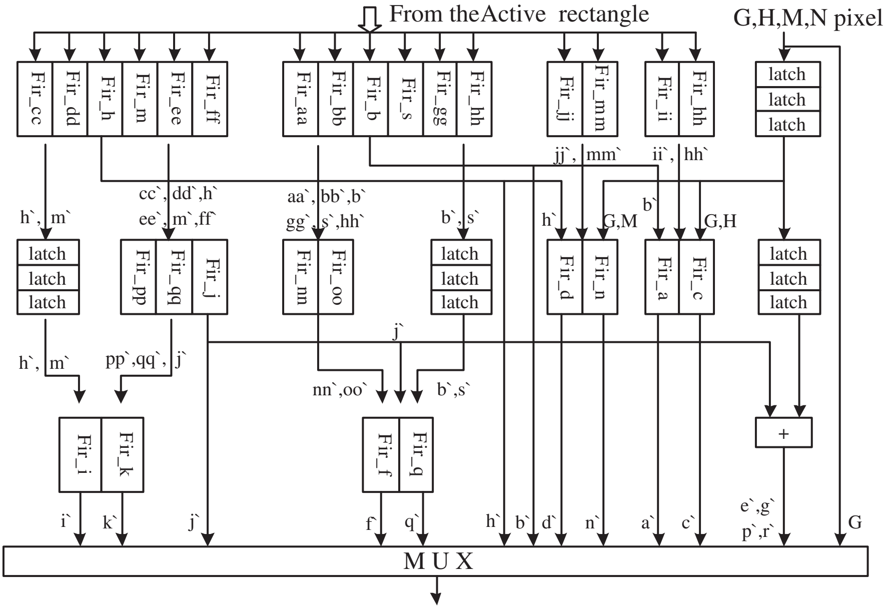


Fig. 3. The ALU

the current processing. And the other is the passive pixel register (P register) which stores the data for the later processing. Active rectangle signed by the shadow area in Fig. 2 has 6 × 6 active pixel registers used to hold the MB samples for the current calculation of 1/2 or 1/4 interpolation. Passive rectangle composed of $(N - 1) \times 6$ passive pixel registers is used to buffer the pixel data for later calculation. The data transfer in three ways: upwards, downwards and to the left.

- *To the left*: one column with $N + 5$ pixels is shifted into the most right side of the array through the set of input registers, at the same time each pixel in the array is shifted one position to the left.
- *Downwards*: pixels are shifted downward (in Fig. 2) one position per cycle.
- *Upwards*: pixels are shifted upward (in Fig. 2) one position per cycle.

For a $N \times M$ block partition,the pixels transfer scheme is presented as below:

```
6 operations of "to the left"; (Initial data in the buffer)
For(I=0; I<M/2; I++)begin
    N-1 operations of"downwards";   (for an even line)
    1 operations of "to the left";
    N-1 operations of "upwards";    (for an odd line)
    if(i<M/2-1) operations of "to the left";
end
```

So total cycles needed by interpolating a partition of N x M is:

$$cycle_{N\times M} = 6 \times cycle_{left} + (N - 1 + cycle_{left}) \times (M - 1) + cycle_{delay} \quad (1)$$

Where $cycle_{left}$ represents the cycles of the operation *to the left*, and $cycle_{delay}$ represents the cycles delayed by ALU showed in Fig. 3.

3.2 The ALU

The ALU (in Fig. 3) is responsible for the calculation of the fractional samples. According section 2, these samples are classified by five cases. The ALU has 29 filters in order to have the parallelization in all conditions of motion vector. Block named Fir_x is the special used filter for processing the fractional sample x. And the architecture of Fir_x is shown in Fig. 4 in which an adder tree is adopted instead of the multiplication. The data for calculating a certain fractional sample are derived directly from the active rectangle in Fig. 2 simultaneously and inputted into the ALU at the same time also. In Fig. 3, Some full samples or half samples must be delayed for matching the latency of filters, such as G for calculating a and h for k. The function of MUX is to select the output pixels according to motion vector. The operation of *clip1* is in the MUX, and applied before pixels are exported.

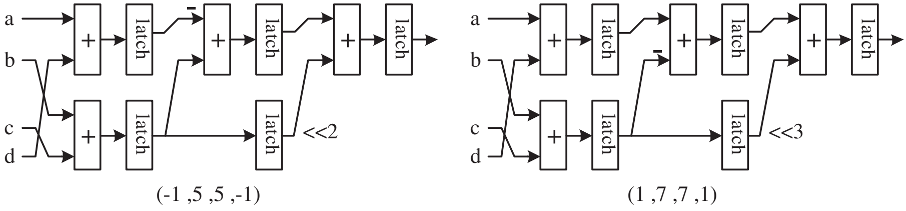


Fig. 4. The Architecture of filter

4 Experimental Results

Since the whole AVS decoder pipeline is based on 8x8 blocks, the 8x8 block partition is adopted in the implementation of the architecture in Fig. 2. However, by control, the implementation can fit all the partitions.

The implementation was described using Verilog-HDL, and synthesized with synopsys tools using 0.25um Standard Cell Library. The total area is about 379715 um^2. The $cycle_{delay}$ of the eq.(1) in the ALU is 9 cycles at most and the $cycle_{left}$ is one cycle in the implementation. So the maximum number of the cycles in 8x8 partition is 71. And the critical path of the architecture is 4.83ns which may satisfy the decoder applications such as SDTV or HDTV.

5 Conclusion

An efficient architecture for MC interpolation of AVS is proposed in this paper. The data transfer scheme of this architecture solves the problem of memory access in MC perfectly. The experimental results show that the proposed architecture can meet the need for the real-time implementation of AVS decoder for SDTV or HDTV.

References

1. http://www.avs.org.cn/en/index.asp
2. L.vos and M.Stegherr: Parameterizable VLSI Architectures for the full-search Block-Matching Algorithm. IEEE Transactions on Circuits and Systems, 36(10):1309–1316, October 1989
3. J.Park, K.Muhammad and K. Roy: High Performance FIR Filter Design Based on Sharing Multiplication. IEEE Transactions on VLSI Systems (TVLSI), Vol.11, Issue: 2, pp: 244–253, April 2003
4. H.Samueli: On the design of optimal equiripple FIR digital filters for data transmission applications. IEEE Trans. Circuits Syst., vol. 35, pp. 1542–1546, Dec 1988
5. N.Sankarayya, K.Roy, and D. Bhattacharya: Algorithms for low power and high speed FIR filter realization using differential coefficients. IEEE Trans. Circuits Syst., vol. 44, pp. 488–497, June 1997
6. JVT:ISO/IEC and ITU-T: Draft ITU-T Recommendation and Final Draft international Standard of Joint Video Specification. Doc.JVT-G050r1, Geneva, Switzerland, May, 2003
7. Nuno Roma, Leonel Sousa: A New Efficient VLSI Architecture for Full Search Block Matching Motion Estimation. VLSI-SOC, pp. 253–264, 2001: Montpellier, France

Fast Fractal Image Encoder Using Non-overlapped Block Classification and Simplified Isometry Testing Scheme

Youngjoon Han[1], Hawik Chung[2], and Hernsoo Hahn[1]

[1] Soongsil University, Seoul, Korea
{young,hahn}@ssu.ac.kr
[2] Kyungbok College, Pocheon city, Gyeonggi-do, Korea
hichung@kyungbok.ac.kr

Abstract. This paper aims at reducing the encoding time of a fractal image encoder. For this purpose, a non-overlapped block classification method and a simplified isometry testing scheme are proposed. The non-overlapped block classification method avoids the repeated classification operations needed for finding the domain blocks having the same type with the range block, by memorizing the classification results of the domain blocks and using them for the overlapped blocks in a new searching area. For reducing the time required for calculating a similarity between blocks, a simplified isometry testing scheme is used. It tests the isometry between a domain block and a range block using only those types of isometry having the similar features with the type of the range block. For speeding up the calculation time, the SOFM neural network is used as the block classifier and the spiral searching scheme is used. The experimental results have shown that the proposed algorithm reduces the encoding time by 50% on average while maintaining the same PSNR and bit rate, compared to the other's recent approaches.

1 Introduction

In multimedia communication, image data takes the most potion of the transmission time. To reduce the image size and thus to enhance the multimedia transmission time, various image coding methods have been introduced, such as vector quantization, subband coding, predictive coding, and fractal coding, etc(see [1]). Among these methods, it is well known that the fractal coding method has a highly compressed rate and there occurs almost no distortion when the compressed image is enlarged. Despite of these attractions, the fractal image encoder have not been actively used in real applications, due to its long encoding time(see [1,2]).

To reduce the encoding time while keeping the advantages of the fractal image encoder, recently many researchers have proposed various encoding algorithms. These efforts have started from Jacquin's work([3]). He proposed a block classification method to reduce the comparison time required for finding similarity in an image. He classified the range and domain blocks into 3 types: shade blocks,

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 207–214, 2004.

edge blocks, and midrange blocks. In shade blocks, the image intensity varies only very little, while in edge blocks a strong change of intensity occurs along a boundary of an object. Midrange blocks have blocks have larger intensity variations than shade blocks. Another method of defining the classes is the archetype classification proposed by Boss and Jacobs([4]). An archetype for a set of codebook blocks are given by the particular codebook block that can best cover all others in the usual least squares sense. Starting out from an arbitrary classification, the archetype for each class is obtained, and then the blocks are reclassified according to the archetype by which they can be covered best. The final set of archetypes becomes a part of the encoder. Jean Cardinal proposed the method of using features vectors and domain classification in the matching process in [5]. It computes feature vectors for ranges and domains, then the vector space is recursively partitioned until each partition contains a sufficiently small number of range and domain. Finally, range/domain comparisons are made inside each partition, storing the best transformation for each range. Despite these efforts, the encoding time still remains as the problem that must be solved.

This paper deals with two problems to reduce the encoding time in fractal coding method based on block classification approach. One is how to efficiently classify the blocks and the other one is how to quickly test the isometry between range and domain blocks. The blocks are categorized into four classes using SOFM(self-organizing feature maps) neural network for fast and accurate classification. Since the SOFM neural network has a fast learning time and a self learning function, it can be efficiently used for classifying images with different characteristics. The similarity between a range block and a domain block is tested in different ways depending on their classes. For example, if the type of a range block is flat (that is, if there is no directional feature in the block), then all types of isometry are tested. Otherwise, only the types of the isometry associated with the type of the block are tested. The performance of the proposed algorithm has been tested with the images having different features.

2 Non-overlapped Block Classification

2.1 Conventional Block Classification Scheme

Block classification methods are provided for reducing the number of the similarity tests between a domain block and range block by practicing the test only when they have the same block type. That is, the type of each domain block should be determined to test whether it has the same type with the given range block before calculating the parameters of the affine transformation. The conventional block classification scheme determines the types of the domain blocks whenever a new range block is given. Therefore the number of classifications of the domain blocks becomes very large.

For example, let's consider an image whose size is M × M pixels and the sizes of a range block and a domain block are given as R × R and D × D, respectively. Then the numbers of range blocks N_{R} and domain blocks N_{D} are defined as follows, when the blocks are not overlapping.

$$N_{\mathrm{R}} = \left(\frac{\mathrm{M}}{\mathrm{R}}\right)^2, N_{\mathrm{D}} = \left(\frac{\mathrm{M}}{\mathrm{D}}\right)^2 \tag{1}$$

If the domain blocks are overlapped by shifting the position by one pixel, then the number of the domain blocks (N_{DS}) can be defined as follows.

$$N_{DS} = (\mathrm{M} - \mathrm{D}) \times (\mathrm{M} - \mathrm{D}) = (\mathrm{M} - \mathrm{D})^2 \tag{2}$$

Then the number of comparisons N_C to find the domain block corresponding to a given range block can be defined by the following equation.

$$N_C = N_R \times N_{DS} = \mathrm{M}^2 \frac{(\mathrm{M} - \mathrm{D})^2}{\mathrm{R}^2} \tag{3}$$

Since D ≪ M, N_C can be simplified as follows.

$$N_C \cong \frac{\mathrm{M}^4}{\mathrm{R}^2} \tag{4}$$

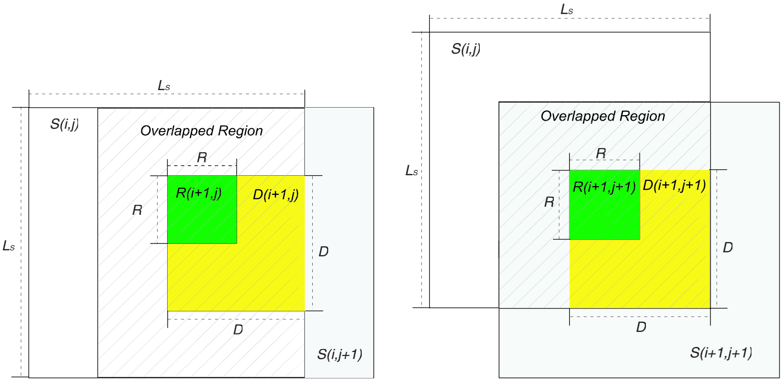
Eq.(4) shows that the image should be segmented into small subimages and the size of range block should be large, to reduce the number of comparisons. Although the equation says that N_C becomes the minimum when M equals to R, that is not effective obviously.

2.2 Non-overlapped Block Classification Scheme

Differently from the conventional method, the proposed scheme classifies only those domain blocks located along the spiral trajectory starting from the given range block. It is invented based on the observation that there is a higher possibility that the domain blocks locating near the range block may have higher similarities than others. Therefore, the searching area for a given range block $R\,(i, j)$ is restricted by $S\,(i, j)$, as shown in Fig. 1. The size of $S\,(i, j)$ is $L_S \times L_S$ and it is determined by considering the number of iterations.

Figure 1 shows how the first domain block $D(i+1, j)$ to be classified is defined for a given range block $R(i+1, j)$. Starting from $D(i+1, j)$, the domain blocks defined along a spiral trajectory are classified and the classification results (the types of the domain blocks) are stored in a reference memory. If a new range block to be referenced is $R(i, j)$ next to $R(i + 1, j)$, then the corresponding searching area $S(i+1, j)$ will be changed to $S(i, j)$, as shown in Fig. 1. Since the new range block is next to the old range block, most of the searching area of the new range block is overlapped with that of the old range block. The overlapped portion in the new searching area is shaded in the figure. Since the types of the domain blocks in the shaded area of $S(i, j)$ are stored in the reference memory, the classification process for those domain blocks located in the overlapped area can be replaced by a memory reading operation.

From Fig. 1(a), the size of the overlapped area can be defined as $(L_S - \mathrm{R}) \times \mathrm{L_S}$. If the new range block is $R(i + 1, j + 1)$ as shown in Fig. 1(b), the overlapped area is defined as ${L_S}^2 - R^2$. Therefore, the overlapped searching region where



(a) Horizontally overlapped area (b) Diagonally overlapped area

Fig. 1. Searching areas of two neighboring range blocks

the classification will not be processed can be defined by the following equation, if the domain blocks in the searching area $S(0, 0)$ is fully classified.

$$O_{XY} = \left| \begin{array}{ll} (L_S - \mathrm{R}) \times \mathrm{L_S} & i = 0 \text{ or } j = 0 \text{ in } R\,(i, j) \\ {L_S}^2 - \mathrm{R}^2 & i \neq 0 \text{ and } j \neq 0 \text{ in } R(i, j) \end{array} \right| \tag{5}$$

By avoiding the classification operations in this overlapped area, the number of classification operations for the domain blocks can be reduced up to 90.87% in theory.

2.3 Block Classifier Using SOFM Neural Network

In this paper, the range and domain blocks are classified into four types, as done in [6,7]: flat, mixed, vertical/horizontal, and diagonal. 'Flat' is defined when there is almost no intensity variation in the block. 'Mixed' is defined when there is a certain amount of intensity variation without any pattern. 'Vertical'/'horizontal' is the type where the variation occurs vertically or horizontally. 'Diagonal' is the type where there is variation in diagonal direction. The sizes of the domain and range blocks are 8×8 and 4×4 pixels respectively.

To speed up the block classification process, the SOFM neural network based classifier is used which is trained by the Kohonen's learning algorithm. It operates in two modes, one for range block classification and one for domain block classification. When operating in a domain block classification mode, it uses the 64 input nodes corresponding to 64 pixels in the domain block, as shown in Fig. 2.

3 An Improved Isometry Testing Scheme

The conventional methods determine the similarity between a domain block and a range block based on the coefficients of the transformation. To reduce the

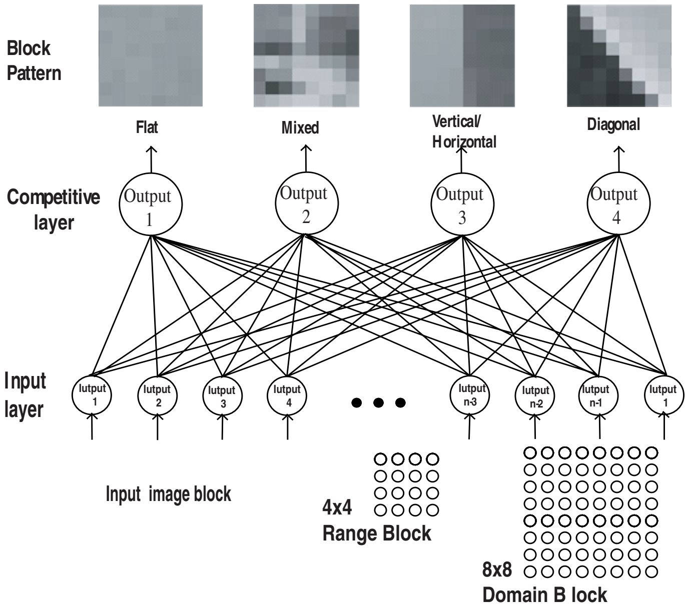


Fig. 2. SOFM neural network based block classifier

calculation time of the transformation, Jacquin proposed a simplified calculation method which considers a domain block to be isometric with the range block if their rotational relationship satisfies one of the 8 types of isometry(see [7, 8]).That is, each type of the isometry represents the degree of rotation between the domain block and the range block. For finding the best matching type of the isometry for the given range block, the domain block is rotated 8 times with reference to the 8 types of the isomtry. Therefore 8 comparisons per one domain block are required to determine the isomorphism between a domain block and a range block.

To reduce this number of comparisons for testing the similarity, this paper proposes an improved isometry testing scheme which tests the types of the isometry which are related with the type of the block. For example, if the type of the range block is 'flat', since there is no intensity variation in the block, there is no need to test how much a block is rotated. The test of the non-rotated type of isometry is enough. By the similar reason, if the type of the range block is 'mixed', all 8 types of isometry should be tested since there exists an equal possibility for all 8 types of the isometry. If the type of the range block is 'horizontal/vertical' or 'diagonal' types, then only those types of isometry having the similar features with the range block should be tested. The types of isomtry that should be tested with reference to the type of the range block are summarized in Table 1.

To see how much time is saved by the proposed method, let's assume t_m be the time required for testing one type of isometry. Then the total time to test

Table 1. Types of isometry to be tested with reference to the types of the block

Type of block	Number of isometry tests	Types of isometry
Flat	1	Original image
Mixed	8	8 directions
Horizontal/Vertical	3	Original image, Rotated image about X & Y axis
Diagonal	3	Original image, Rotated image about X=Y & X=-Y axis

all 8 types of isomtry for n blocks, represented by T_8, becomes $8 \times n \times t_m$. This case can be considered as all blocks have the 'mixed' type. If the types of blocks are spread and A, B, C, and D are the proportions of blocks pertained to 'flat', 'mixed', 'vertical/horizontal', and 'diagonal' types, respectively, then the total time T_P to test the isometry becomes as follows.

$$T_p = t_m \times n \times \left(\frac{A}{100} + \frac{8B}{100} + \frac{3C}{100} + \frac{3D}{100} \right) \tag{6}$$

Then the ratio of T_P with reference to T_8 can be expressed by the following equation.

$$T_p/T_8 = \frac{A + 8B + 3C + 3D}{800} \tag{7}$$

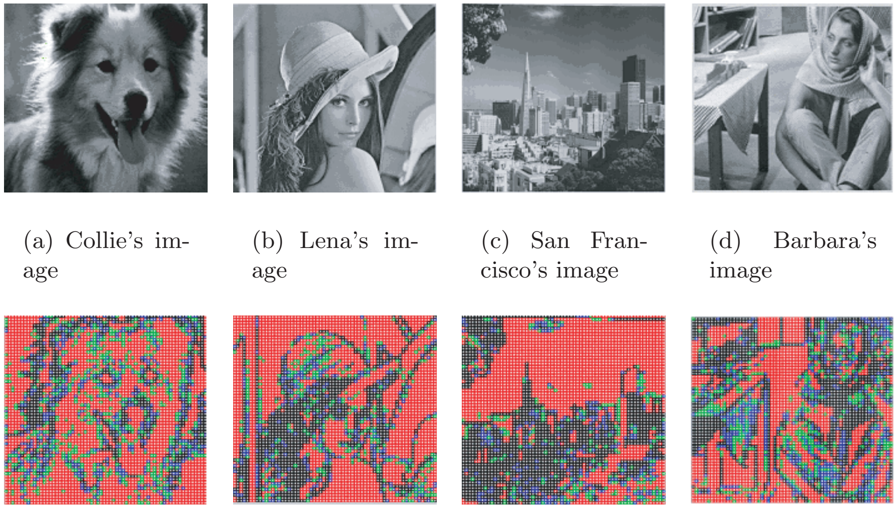
As summarized in Table 2, the largest proportion of the images, most popularly used for evaluating the performance of the image coding methods, contain the 'flat' typed blocks which require a test of only one type of isometry, resulting in a significant reduction of the total test time. The table shows that the proposed algorithm saves the similarity test time by 60% on an average.

Table 2. Proportions of the block types in the images and the ratio of T_P to T_8

Image	Flat	Mixed	Vertical Horizontal	Diagonal	T_P(n=4096)	T_P/T_8(n=4096)
Collie	54.7%	16.8%	16.1%	12.4%	$11247.6t_m$	34.3%
Lena	55.2%	27.0%	9.4%	8.4%	$13271.7t_m$	40.5%
San Francisco	50.0%	36.0%	6.4%	7.7%	$15577.0t_m$	47.5%
Babara	40.6%	33.1%	12.2%	14.1%	$15740.9t_m$	48.0%

4 Experiments

The proposed algorithm has been implemented using Visual C++ in Pentium IV PC. It was evaluated in terms of the decoded image quality (PSNR), encoding time, and compression rate (bpp), and compared to the algorithms of Bansley and Jacquin. For the purpose of comparisons, four images are selected: Collie's image having a lot of high frequency components, Lena's image having



(a) Collie's image

(b) Lena's image

(c) San Francisco's image

(d) Barbara's image

(e) Classifying (a)

(f) Classifying (b)

(g) Classifying (c)

(h) Classifying (d)

Fig. 3. Test image and their classified images

Table 3. Experimental results of the proposed algorithm

Image	Coding Method	Compression time(sec)	Decoded Quality(PSNR[dB])	Compression Rate(bpp)
Collie's	Barnsely	1363.0	31.90	2.91
	Jacquin	22.3	29.01	2.95
	Proposed	11.4	28.98	2.94
Lena's	Barnsely	1382.0	32.09	2.89
	Jacquin	20.5	29.28	2.91
	Proposed	10.1	29.29	2.92
San Francisco's	Barnsely	1275.0	31.10	2.97
	Jacquin	18.2	29.89	2.95
	Proposed	9.5	29.30	2.99
Barbara's	Barnsely	1305.0	31.90	2.93
	Jacquin	21.9	29.50	2.95
	Proposed	10.9	29.58	2.94

a lot of low frequency component, and San Francisco's image having both high and low frequency components in similar portions. Each of them has a resolution of 256 × 256 pixels. The encoding time includes the time spent in the SOFM neural network to classify the blocks into 4 types and the time spent for testing the isometry.

The test results are summarized in Fig. 3 and Table 3. Figure 3 shows the results of the classification of blocks using the SOFM neural network and Table 3 compares the performance of the proposed algorithm to the other references. The results show that the proposed algorithm reduces approximately 50% of the encoding time while keeping the PSNR of 30dB on average.

5 Conclusion

This paper proposed a fast encoding algorithm for the fractal image encoder using a non-overlapped block classification scheme and a simplified isometry testing scheme. The non-overlapped block classification scheme eliminated overlapped classifications of domain blocks by storing the classification results of the domain blocks in the reference memory and reading the type from the memory when the classified block is to be classified once more. The simplified isometry testing scheme reduces the number of types of isometry used for testing the similarity between a range block and a domain block, by selecting only those types of isometry associated with the type of block. The SOFM neural network based classifier and the spiral searching scheme also contributed for reducing the encoding time further. As shown in the experimental results, the proposed algorithm reduces the classification time about 50% without degrading the PSNR and compression rate.

References

1. Sayood, K.,: Introduction to Data Compression. Academic Press CA. Inc.
2. Barnsely, M.,Jacquin, A.: Application of recurrent iterated function systems to images. Visual communications and image processing, Vol. 1001. (1988) 121–132
3. Jacquin, A.: mage Coding based on a fractal theory of iterated contractive image transformations. IEEE Transaction on Image Process, Vol. 1. (1992) 18–30
4. Boss, R., Jacobs, E.: Archetype classification in an iterated transformation image compression algorithm. Fractal Image Compression – Theory and Application, Springer-Verlag, New York (1994)
5. Cardinal, J.: Fast Fractal Compression of Greyscale Images. IEEE Transaction on Image Processing, Vol. 10. (2001) 159–163
6. Jacobs, E., Fisher, Y.: Image Compression: A Study of the iterated transform method. Signal Process, Vol. 29. (1992) 251–263
7. Wohlberg, B., Jager, G.: A Review of the Fractal Image Coding Literature, IEEE Transactions on Image Processing, Vol. 8. (1997)
8. Li, Y.: An Efficient Fractal Image Compression Method. Proceedings of the 1997 IEEE International Conference on Systems, man, and Cybernetics, (1997) 4204–4206

A Fast Downsizing Video Transcoder Based on H.264/AVC Standard

Chih-Hung Li, Chung-Neng Wang, and Tihao Chiang

National Chiao Tung University, HsinChu, 30050, Taiwan
tchiang@mail.nctu.edu.tw

Abstract. This paper presents a novel spatial transcoding technique for H.264/AVC based video applications. With superior coding performance, H.264 will be used in a variety of multimedia applications. For various video applications, spatial transcoding techniques can facilitate compact storage of video sequences with specified resolution and bitstream format. Under demands, we can convert the archival bitstreams to another bitstream with different bitrates and different resolution before transmission. For H.264 video transmission that needs spatial resolution conversion, we propose a bottom-up motion vector re-estimation, rapid rate-distortion (R-D) optimized mode decision, and adaptive motion search range to speed up the spatial transcoding and retain visual quality of transcoded video. The results show that the fast transcoder has the R-D performance close to or better than R-D re-encoding algorithm, which has the highest complexity and best coding efficiency.

1 Introduction

In many multimedia applications like video on demand or video archive, video content is often stored with specified spatial resolution and compressed format to minimize storage. For compact storage of video data, a video coding standard H.264/AVC [1] has been used. H.264 with many advanced coding technologies can achieve up to approximately 50% bit saving for similar perceptual quality as compared to other existing standards H.261/3 and MPEG-1/2/4. In addition, H.264 employs Rate-Distortion Optimization (RDO) that exhaustively searches for the best coding mode to maximize picture quality of decoded video at the given bitrates, which dramatically increases computational load of coding processes [2]. Thus, fast algorithms are demanded for H.264 based video applications.

For applications including picture-in-picture presentation, TV wall and video browsing with archival bitstreams, transcoding techniques are required to fit channel conditions and receivers' terminal capabilities. Spatial transcoding techniques can facilitate compact storage of video sequences with the specified resolution and bitstream format. Under service request, we can convert the archived H.264 bitstream to another bitstream with different bitrates and different resolution before transmission. For real-time transmission that needs downscaled video sequences, the high complexity inherent from H.264 standard shall be addressed.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 215–223, 2004.

Some studies have provided fast spatial transcoders for existing standards including MPEG1/2/4 and H.261/3 [3]-[9]. For a rapid spatial transcoder based on H.264, Sun et al. [9] estimated the motion vectors required to code the downscaled video sequences by directly extracting the motion vectors from the archival bitstreams. The existing methods focus on the reduction of computationally intensive motion estimation. However, to maximize the coding efficiency of the transcoder, H.264 RDO is required. Since H.264 RDO takes about 60% of overall encoding time, further investigation is needed for realizing fast transcoding that can retain high coding efficiency.

We present a novel and fast downscaling transcoding scheme based on the retrieved information including texture data, motion vectors and R-D optimized coding modes from the input bitstreams with video sequences of larger spatial resolution. For texture extraction, existing spatial transcoders have adopted the DCT-domain downscaling methods [3]-[5], which combine DCT-domain motion compensation and DCT-domain block synthesizing to speed up the extraction process. To fit into H.264/AVC based spatial transcoding, since the spatial-domain tools consisting of the sub-pixel interpolation filter, deblocking filter and intra prediction are used for H.264 decoding, we convert the frame resolution in pixel domain.

Motion vector re-estimation (MVR) is sped up in spatial transcoders based on the existing video specifications [7]-[9] In H.264, considering seven inter prediction modes, we present a new MVR approach, called as bottom-up MVR (BUMVR). BUMVR can reduce computation power based on the correlation of motion vectors for different block sizes. In addition, to accelerate the RDO mode decision that can retain high picture quality, we eliminate some inter or intra modes that have lower possibility of being selected as the best mode for coding each block. Experiment results show that with the extracted information, our method can outperform the existing methods in terms of transcoding frame rates and R-D performance. For a downscaling video transcoding, R-D performance of the proposed algorithm is close to or even better than that of R-D re-encoding algorithm, which has the highest complexity and the best coding efficiency. Our novel downscaling video transcoder has about 8 times speedup on average over R-D re-encoding spatial transcoding.

2 Spatial Transcoding Based on H.264/AVC

For H.264, two spatial transcoders including rate-distortion re-encoding (RDRE) and top-down MVR (TDMVR) are introduced for performance comparison. RDRE fully decompresses the incoming bitstream, subsamples the frame and then compresses the downscaled frames into a new bitstream. RDRE consists of a decoder, a spatial-domain down-sampling module and an encoder, which are aligned in a cascaded manner. RDRE consumes a great amount of computation power, which limits its practical application. For transcoding, the re-encoding process that re-quantizes the reconstructed coefficients will introduce additional quality loss to the transcoded video sequences. Thus, it is challenging to build a fast downscaling video transcoder with proper visual quality.

To reduce computational load of the downscaling video transcoding, some fast MVR algorithms have been proposed to estimate the motion vectors of the downsized frames with the motion vectors of the frames within archival bitstreams [7]-[9]. In H.264, the application of all 7 inter prediction modes is computationally expensive. To speed up the transcoding, TDMVR [10] that derives the motion vectors with the median operation in a large-to-small manner used only 4 inter modes with the block size larger than 8×8 for downsized frames under the assumption that using the block sizes smaller than 8 × 8 for MVR can only provide few coding gain at high bit rates. However, to apply TDMVR into the downsized frames may eliminate local motion information of the pre-coded frames. In addition, TDMVR has not resolved the RDO mode decision complexity.

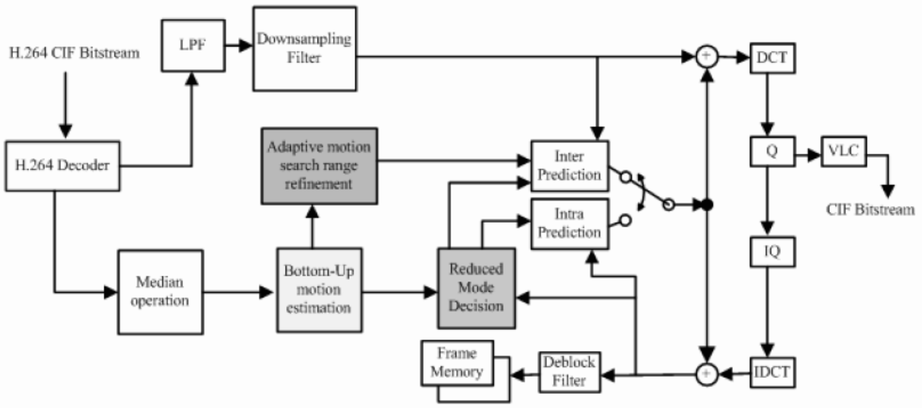


Fig. 1. Architecture of proposed spatial video transcoder.

3 Fast Spatial Transcoding Based on H.264/AVC

Figure 1 shows the architecture of a H.264 based fast downscaling video transcoder. The fast spatial transcoding is based on the reuse of the content extracted from the input bitstreams. As compared with RDRE, we have added three new modules including BUMVR, rapid mode decision and adaptive motion search range. In Figure 1, for the low-pass filter (LPF) and the downsampling filter, we adopt a 4-to-1 downsampling scheme with an average operation. Other complicated re-sampling methods and low-pass filters can be applied for frame size conversion with extra computation.

3.1 Bottom-Up Motion Vector Re-estimation (BUMVR)

TDMVR will lose lots of motion information due to its larger initial block. Therefore, we propose BUMVR to preserve more motion information with its smaller initial block. In addition, TDMVR that predicts the motion vectors of other modes from spatial neighboring motion vectors may not find the best reference when motion vector fields are not homogeneous. BUMVR can predict the motion vectors from the sub-blocks since all the blocks of different sizes locate at the same macroblock, which menas that the motion vectors of the subblocks within a MB have strong correlation.

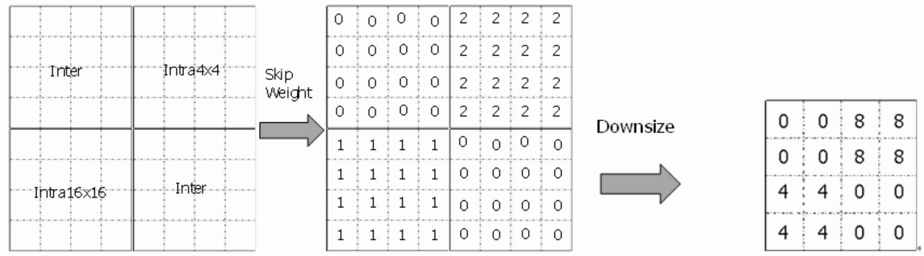
With the motion information remaining in the input bitstreams, BUMVR can enhance the precision of the motion vectors for blocks with large sizes. The enhancement is based on merging the motion vectors of smaller blocks in a bottom-up manner. Initially, we derive the motion vector of each 4×4 block in the downsized frame with the median value of the motion offsets that the corresponding four 4×4 blocks within the archival frame have. For computation reduction, the simple median operation instead of the operation in TDMVR [10] is used. Based on high correlation between 4×4 blocks in a MB, we further combine the motion vectors to obtain the motion vectors of other modes with block sizes larger than 4×4. The combination approaches for various modes are summarized at Table 1. In the spatial transcoder, the motion vectors of Inter 8×4 and Inter 4×8 can be derived as the averaged value of two motion vectors from the corresponding pair of 4×4 blocks. Motion vectors of Inter 8×8 and other modes with block sizes larger than 8×8 are set as the median value of motion vectors from a group of 4×4 blocks. Thus, with the derived motion vectors of 4×4 blocks and the simple combination methods, BUMVR can rebuild all motion vectors that are close to the motion vectors in the input bitstreams for inter prediction in transcoding.

Table 1. Bottom-up merging of motion vectors

Inter mode	Estimated motion vectors
1	$mv''_{1} = Median\{mv'_{1}, mv'_{2}, ..., mv'_{16}\}$
2	$mv''_{21} = Median\{mv'_{1}, mv'_{2}, ..., mv'_{8}\}$ $mv''_{22} = Median\{mv'_{9}, mv'_{10}, ..., mv'_{16}\}$
3	$mv''_{31} = Median\{mv'_{i} \mid i = 1, 2, 3, 4, 9, 10, 11, 12\}$ $mv''_{32} = Median\{mv'_{i} \mid i = 5, 6, 7, 8, 13, 14, 15, 16\}$
4	$mv''_{4i} = Median\{mv'_{4i-3}, mv'_{4i-2}, mv'_{4i-1}, mv'_{4i}\}$
5	$mv''_{5i} = Average\{mv'_{2i-1}, mv'_{2i}\}$
6	$mv''_{6i} = Average\{mv'_{i}, mv'_{i+2}\}$
7	$mv''_{7i} = mv'_{i}$

3.2 Rapid Mode Decision

To balance the computation cost of mode decision and the reconstructed picture quality, the RDO coding modes of the archival frames are referred to provide the content properties of the source video sequences for R-D optimized transcoding. Thus, the rapid mode decision is based on removal of modes that do not facilitate RDO coding. For example, if a block and the neighboring blocks within a frame are all coded as intra mode, the block of down-sampled frame should be coded as intra mode. Thus, with the extracted information, we can speed up the mode decision by eliminating some inter modes or intra modes that have lower probability of being the best coding mode of the handling blocks. The skip mode that has less computational load is not discussed here.



0	0	0	0	2	2	2	2
0	0	0	0	2	2	2	2
0	0	0	0	2	2	2	2
0	0	0	0	2	2	2	2
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0

0	0	8	8
0	0	8	8
4	4	0	0
4	4	0	0

Fig. 2. Intra factor map of original frame and downsized frame. The grid in the figure indicates the boundary of a 4 × 4 block.

Intra Mode Decision. Intra mode decision takes more computations than inter mode selection. The number of intra mode combinations for luma and chroma blocks in each MB equals to $M_8(16 \times M_4 + M_{16})$ where M_8,M_4 and M_{16} represent the number of possible modes for 8 × 8 chroma blocks, 4 × 4 and 16 × 16 luma blocks respectively. Thus, each MB has to take 592 R-D calculations to derive the best mode with the lowest R-D cost. The intra factor is used to denote the possibility to choose each type of intra modes as the best mode for transcoding in terms of R-D performance. The intra mode reduction consists of two steps including creating intra factor maps and re-encoding the downsized frames. Figure 2 illustrates the concept of intra factor maps.

The intra factor map of the pre-coded frame is built up by assigning the intra factor values for different modes. The factors of Intra4 × 4, Intra16 × 16 , and Inter modes are set as 2, 1 and 0 respectively. The intra factor map of the down-sampled frame is obtained by summing up the intra factor values of the corresponding blocks within the pre-coded frames. With the intra factor map of the down-sampled frame, we only need to justify all intra modes for the blocks with the intra factor larger than a specified threshold in RDO calculation. Table 2 shows the thresholds found empirically. As the QP increases, we increase the threshold for Intra4 × 4 mode and decease the threshold of Intra16 × 16 based on the observations that Intra4 × 4 mode has higher probability to become the best mode at higher bit rates and Intra16 × 16 mode has higher probability to be the best mode at lower bit rates. Thus, by decreasing the percentage of the macroblocks that have to evaluate intra modes in RDO coding, we can speed up the intra mode decision of H.264 based spatial transcoding.

Table 2. Thresholds of intra mode decision

QP range	Intra4 × 4	Intra16 × 16
$\text{QP} \leq 35$	48	48
$35 < \text{QP} \leq 40$	64	48
$40 < \text{QP} \leq 45$	80	32
$\text{QP} > 45$	96	32

Inter Mode Decision. For further speedup, we investigate the maximum percentage of the blocks that have to verify inter modes. Since the possible combinations of the inter modes for transcoding can be inherent from the coding modes in the input bitstream, we can reduce the complexity of inter mode decision for the H.264 based transcoder. The inter mode decision is applied when the motion vectors of subblocks within a MB are inconsistent. The motion vector consistency is examined based on the difference between the motion vectors of subblocks in a block. Thus, the motion vector consistency is measured by

$$D_{i,j} = \mid mv_i - mv_j \mid < (Thr + \sqrt{QP_r - QP_o}) \text{ , for } i, j = 1, 2, 3, 4 \tag{1}$$

Where QP_o indicates the quantization step size of the incoming bitstream and QP_r indicates the re-quantization step size in transcoding. $\sqrt{QP_o - QP_r}$ shows that as QP_r increases or the output bitrates decreases, larger block partitions are most probably chosen as the best mode and smaller block partitions will be excluded from RDO computation. Thus, for 8 × 8 blocks, as the difference is less than a specified threshold $Thr = Threshold\_b4$, three modes including Inter4 × 4, Inter4 × 8 and Inter8 × 4 modes will be removed from RDO calculation. For a MB, if the difference is less than $Thr = Threshold\_b8$, we eliminate three modes covering Inter8 × 8 , Inter16 × 8 and Inter8 × 16 modes. In our simulations, $Threshold\_b8$ is 0. In addition, the observations on the R-D performance show that $Threshold\_b4$ could be derived by

$$Threshold\_b4 = 1 + \frac{SumDifference}{N \times ScaleFactor} \tag{2}$$

In MVR with the 4-to-1 median operation, we check six relative values of motion vectors. If any pair of motion vectors has different values, we add one to N and sum up the difference to $SumDifference$. $ScaleFactor$ is the reciprocal of geometry ratio used to scale down frames. $Threshold\_b4$ is updated frame by frame.

3.3 Adaptive Motion Search Range

To improve the reconstructed quality and maintain fast transcoding process simultaneously, an adaptive reduction of motion search range is used to refine the motion vectors. Based on the correlation between the predicted motion vectors from the three surrounding blocks and re-estimated motion vector of the current block found via BUMVR, the search range for motion vector refinement is adapted by

$$SearchRange = 1 + \mid mv_p - mv_r \mid \tag{3}$$

In addition, the central location of motion search is moved to the middle position of predicted and re-estimated motion vectors.

$$SearchCenter = \frac{mv_p + mv_r}{2} \tag{4}$$

Consequently, we increase the search area for high motion objects and decrease the area for slow motion objects, which can facilitate the fast motion vector refinement by reducing the amount of search points. Experiment results show that the averaged search range is about 3 for refinement. Thus, the search points per MB are cut down from 1089 to 49. In addition, the R-D performance of adaptive search range is better than that of fixed search range. The rationale is the motion vector refinement within the adaptive search range has high possibility to detect the motion vector candidates with the globally minimal R-D cost.

4 Experimental Results

Performance comparison on a variety of spatial transcoders is based on the following factors including the MVR, the mode decision, the video sequences of CIF (352x288) resolution, the computation time and R-D performance. The simulation platform is Windows 2000 running on an AMD XP2500+ machine with 768MB RAM.

Experimental results in Figure 3 show that BUMVR can retain superior R-D performance than that of TDMVR. In addition, in some cases, the performance of BUMVR is close to the performance of RDRE that takes a great amount of computation power. Observations on the transcoding frame rates in Figure 4 show that the reduction of inter mode decision will speed up the transcoding by averaged 8 times with PSNR degradation about 0.2 to 2.0 dB for different sequences and bit rates. In addition, the results show that adaptive search range can improve the transcoding performance by reducing the total number of search points to about 5% of the search area used in full-search motion estimation. To balance the transcoding speed and the visual quality of reconstructed video, the proper reduction of the mode decision is required.

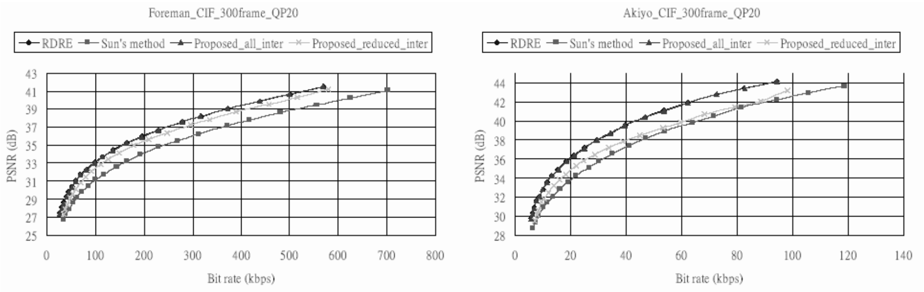


Fig. 3. R-D performance of various downscaling video transcoders

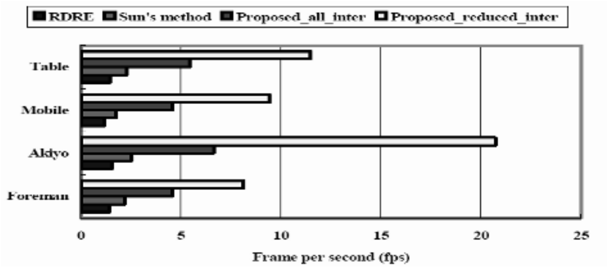


Fig. 4. Frame rates of various transcoders.

5 Conclusions

We have developed a fast downsizing video transcoder consisting of the BUMVR, rapid mode decision and adaptive search range. The BUMVR utilizes the motion information in the bottom-up merging process to obtain motion vectors of all inter mode partitions, which can eliminate the computation load of motion estimation. In addition, by rapid mode decision that reduces the percentage of the RDO calculation based on block characteristics, we can speed up the H.264 based downsizing transcoding. The adaptive motion search range refines the motion vectors derived from BUMVR method in a small checking area. With the three fast transcoding methods, the proposed fast downscaling video transcoder is better than existing methods in terms of R-D performance and transcoding frame rates.

References

1. “Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC),” in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050, 2003
2. T. Wiegand, G.J Sullivan, G. Bjntegaard and A. Luthra, “Overview of the H.264/AVC video coding standard,” IEEE Trans. Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560-576, Jul. 2003

3. K. Seo and J. Kim, “Fast motion vector refinement for MPEG-1 to MPEG-4 transcoding with spatial down-sampling in DCT domain,” in Proc. ICIP’01, vol. 1, Oct. 2001, pp.469–472
4. N. Kim, Y. Kim, G. Kwon and S. Ko, “A fast DCT domain downsampling technique for video transcoder,” in Proc. ICCE’03, June 2003, pp. 36–37
5. Y. Lee, C. W. Lin and Y. W. Chen, “Computation reduction in cascaded DCT-domain video downscaling transcoding,” in Proc. ISCAS ’03, vol. 4, May 2003, pp.860–863
6. J. Xin, M.T. Sun, B. S. Choi and K.W. Chun, “An HDTV-to-SDTV spatial transcoder,” IEEE Trans. Circuits and Systems for Video Technology, vol. 12, no. 11, pp. 998–1008, Nov. 2002
7. M.J. Chen, M.C. Chu and S.Y. Lo, “Motion vector composition algorithm for spatial scalability in compressed video,” IEEE Trans. Consumer Electronics, vol. 47, no. 3, pp. 319–325, Aug. 2001
8. B. Shen, I. K. Sethi and B. Vasudev, “Adaptive motion-vector resampling for compressed video downscaling,” IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 929–936, Sep. 1999
9. H. Sun and Y.-P. Tan, “Arbitrary downsizing video transcoding using H.26L standard,” in Proc. ICIP’03, Sept., 2003, pp.173–176

Temporal Error Concealment with Block Boundary Smoothing

Woong Il Choi and Byeungwoo Jeon

Sungkyunkwan University, 300 Chunchun-Dong Jangan-Gu Suwon, Korea,
creata@ece.skku.ac.kr, bjeon@yurim.skku.ac.kr

Abstract. When a block is lost due to a transmission error in compressed bitstream, conventional temporal error concealment schemes try to conceal the lost block by recovering its motion vector based on some matching criteria. Since essential information for deblocking process and residual data are lost, discontinuity occurs at block boundary, and this discontinuity greatly degrades subjective quality of reconstructed video. To enhance the subjective video quality caused by the discontinuity, we propose a new temporal error concealment (TEC) technique capable of block boundary smoothing. In the proposed scheme, the extended boundary pixels in reference frame are overlapped with the lost block boundary with some weighting factor. For effective error concealment, the flexible macroblock ordering (FMO) technique in H.264 standard is used. The experimental results show that the proposed method provides enhanced subjective quality especially in homogeneous region.

1 Introduction

The growing need for location-independent access to multimedia services containing video demands techniques for efficient video transmission over wireless network. The coding efficiency and error resilience are the most important features for video services over wireless network due to limited channel bandwidth and error-prone environment. In many cases, however, error resilient coding tools have some redundancy, because they need to utilize additional bits to detect or recover errors which are occurred in video bitstream. This redundancy makes coding efficiency decrease. When the error resilient tools in H.264 standard, for example, redundant slice (RS) and flexible macroblock ordering (FMO) schemes are used in encoding process, the size of bitstream would be increased due to the redundancy [1].

In contrast to error resilient tools, error concealment techniques do not require extra redundancy. Since they perform only in decoder, the encoded bitstream needs not be changed. Besides, it is the most important factor which leads to the greatest improvement in subjective quality. The damaged video scene due to transmission error can not be recovered without error concealment technique. The damage in a reconstructed frame affects following frames due to the motion compensation process. Therefore, we can say that the recovery of damaged

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 224–231, 2004.

frames critically depends on the performance of error concealment method. This is the basic motivation of this paper.

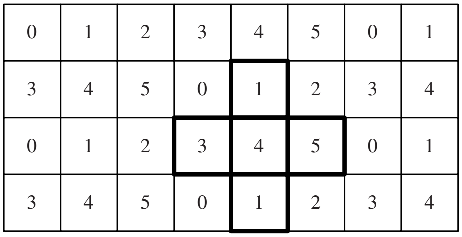
Temporal Error Concealment (TEC) technique conceals a lost block by using the best matched block found in reference frame under some matching criteria, which has been proved to provide good performance in terms of both objective and subjective quality. When conventional TEC is applied for recovering erroneous block, however, it is observed that severe discontinuity sometimes occurs at block boundaries of the concealed block. This is because the in-loop deblocking process cannot be performed in concealed block. To overcome the blocking artifact problem occurred in conventional TEC scheme, we propose a new TEC method by using block boundary smoothing. In simulation result, it is proved that the proposed scheme provide substantial improvement in terms of subjective quality by compensating the discontinuity in lost block boundary. In this paper, we also introduce flexible macroblock ordering technique which assists for TEC scheme to estimate the motion vector of lost block.

2 Flexible Macroblock Ordering (FMO) Scheme in H.264 Standard

Recently released H.264/MPEG-4 AVC standard was designed to have two major features: high-compression efficiency and network friendliness. H.264 standard shows significant improvement in terms of coding efficiency by employing various coding techniques. As a network friendly feature, the Network Abstraction Layer (NAL) in H.264 allows transporting encoded video data over any existing and future network including wireless systems. Besides, various error resilient coding tools such as RS (Redundant Slices), SP (Synchronized Predictive) picture and FMO, assist to handle erroneous bitstreams. These features make H.264 video coding an attractive candidate for video service in mobile environment.

One of reasons that FMO technique was adopted in H.264 is to improve the capability of error concealment. Since TEC schemes utilize the motion vector or pixel values of neighboring blocks to conceal the lost block, the performance of TEC depends on how many correctly decoded neighboring blocks are available. In case of FMO scheme, neighboring blocks used in error concealment can be transmitted in a separate packet by assigning different slice group to each block. If a packet is lost by transmission error or traffic congestion in case of FMO, the lost macroblocks can be easily concealed by using available neighboring blocks in conventional TEC technique.

In case of the dispersed slice structure (it is called FMO mode-1), the slice group of a macroblock is always different from those of neighboring macroblocks as shown in Fig. 1. It means that the neighboring blocks would be sent through different packets. Although the coding efficiency in the dispersed structure would be decreased by preventing the predictive coding using neighboring blocks, the dispersed structure facilitates better utilization of adjacent macroblocks in error concealment of a lost macroblock. In this paper, therefore, we utilize the dispersed slice structure in the proposed TEC scheme.



0	1	2	3	4	5	0	1
3	4	5	0	1	2	3	4
0	1	2	3	4	5	0	1
3	4	5	0	1	2	3	4

Fig. 1. Example of FMO mode 1, dispersed slice structure

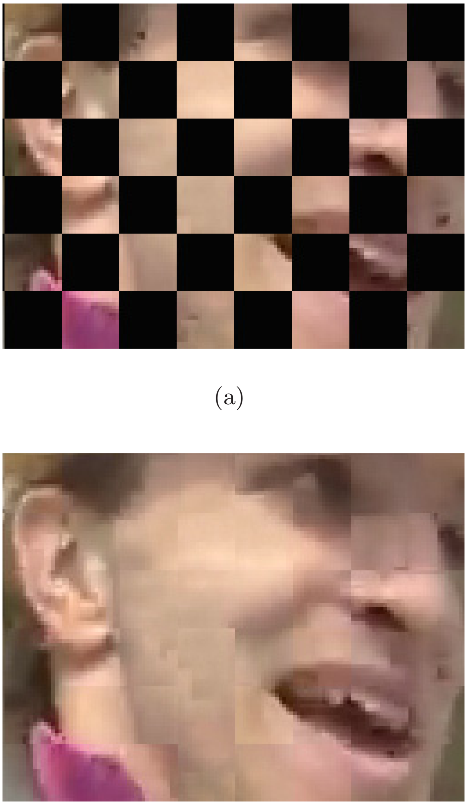
3 Conventional Temporal Error Concealment with Motion Vector Estimation

When some blocks are lost due to packet loss or bit error, temporal error concealment (TEC) techniques attempt to conceal the lost blocks by using temporal correlation within video sequence. TEC techniques require the motion vector of lost block in order to conceal it with a well-matched block in reference frame. To estimate the motion vector of the lost block, some matching criterion is used. Boundary matching algorithm (BMA), one of well-known TEC schemes, tries to find a block which has the minimum distortion between its inner and outer boundaries [2]. Based on smoothness constraint which is the matching criterion of BMA, the motion vector having the minimum distortion will be selected among candidate motion vectors. In BMA, motion vector at neighbor blocks are used for candidate motion vector.

Side matching technique (SMT) measures sum of absolute difference (SAD) of side region between the lost block and reference block which is the block in reference frame indicated by motion vector of the lost block [3]. To recover the motion vector of the lost block, SMT employs the motion search technique like encoder. Since full motion search technique needs substantial amount of computation, fast motion search methods are required in SMT. It has been reported that SMT provides better performance than BMA in terms of PSNR and subjective quality [3]. If the edge of image exists within block boundary, smoothing constraint measurement of BMA fails to find appropriate block among reference blocks due to large distortion in edge region.

4 Error Concealment with Block Boundary Smoothing

Even though conventional TEC schemes find the original motion vector in the lost block, discontinuity occurs at the boundaries of the concealed block. There are two reasons. First, the lost residual data may not be properly recovered although lost block is successfully concealed by TEC scheme. Second, the information needed for deblocking process is lost so that proper deblocking process is not carried out. Note that the H.264 deblocking filter is not post-processing but in-loop processing, and it requires some encoding information, those are the



(a)

(b)

Fig. 2. Example of blocking artifact in conventional TEC: (a) indication of lost block, (b) concealed by SMT

macroblock type information which indicates whether the lost macroblock is inter or intra, quantization parameter (QP), the reference frame number, and the number of transform coefficient required for deciding the block strength value. Since it is not easy to recover both residual data and the information for deblocking process, the blocking artifact is likely to exist in concealed blocks. It causes subjective quality degradation especially in homogeneous region like background.

Figure 2 shows one example of blocking artifact in concealed block when the error concealment is performed by SMT. As depicted in Fig. 2, the black blocks indicate lost blocks due to packet loss or bit error. Since the dispersed slice structure in FMO is used in this example, we can see that the lost macroblocks within the same slice are scattered in reconstructed picture. When the lost blocks are concealed by SMT as shown in Fig 2, it is obvious that the discontinuity at the lost block boundary occurs especially in smooth region like the Foreman's cheek.

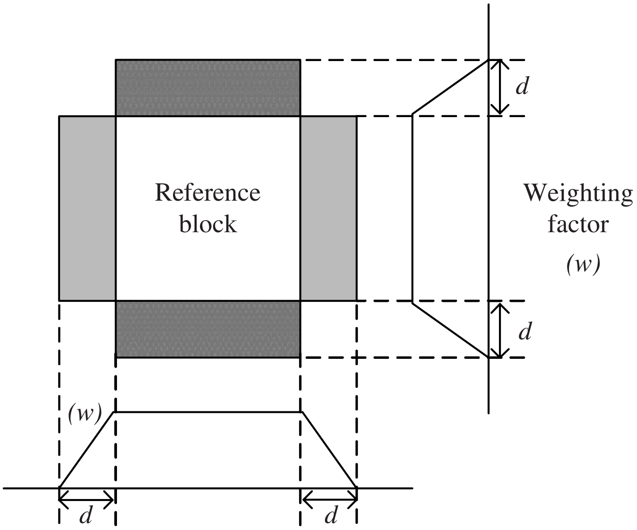


Fig. 3. The proposed TEC scheme with block boundary smoothing: side region in reference block and weighing factor

To overcome the discontinuity around the block boundaries of the lost block, we propose a new error concealment technique with block boundary smoothing. In our proposed scheme, the side region of reference block is also used to contribute to concealment. In TEC schemes, the side region is used to recover the motion vector of the lost block. That is, the reference block having the minimum SAD in side region is decided to conceal the lost block in conventional TEC. Based on this observation, we can see that the discontinuity comes from the gap between each side region. By compensate the difference of side area between lost block and reference block at surrounding the block boundaries, it is expected that the discontinuity can be reduced effectively.

In proposed scheme, therefore, the side region of reference block is overlapped with that of lost block with weighting factor when the reference block is decided to the best matched block for error concealment. Fig. 3 shows the side region of reference block and weighting factor to be used in block overlapping. The lost block, $B_L(x,y)$ is given by

$$B_L(x,y) = w \cdot B_R(x,y) + (1-w) \cdot B_L(x,y) \tag{1}$$

Where, $B_R(x,y)$ is the reference block and w is the weighting factor. The value of weighting factor, w is 1.0 for inside of reference block, and it is linearly decreased as it is far from reference block boundary, as depicted in Fig. 3. By overlapping the side area with linear weighting factor, the discontinuity at the lost block boundary can be smooth out. The strength of smoothness can be decided by the depth of side region, d.

Table 1. PSNR result of each sequence with regard to depth, d (BER=$5 \times 10^{-}4$)

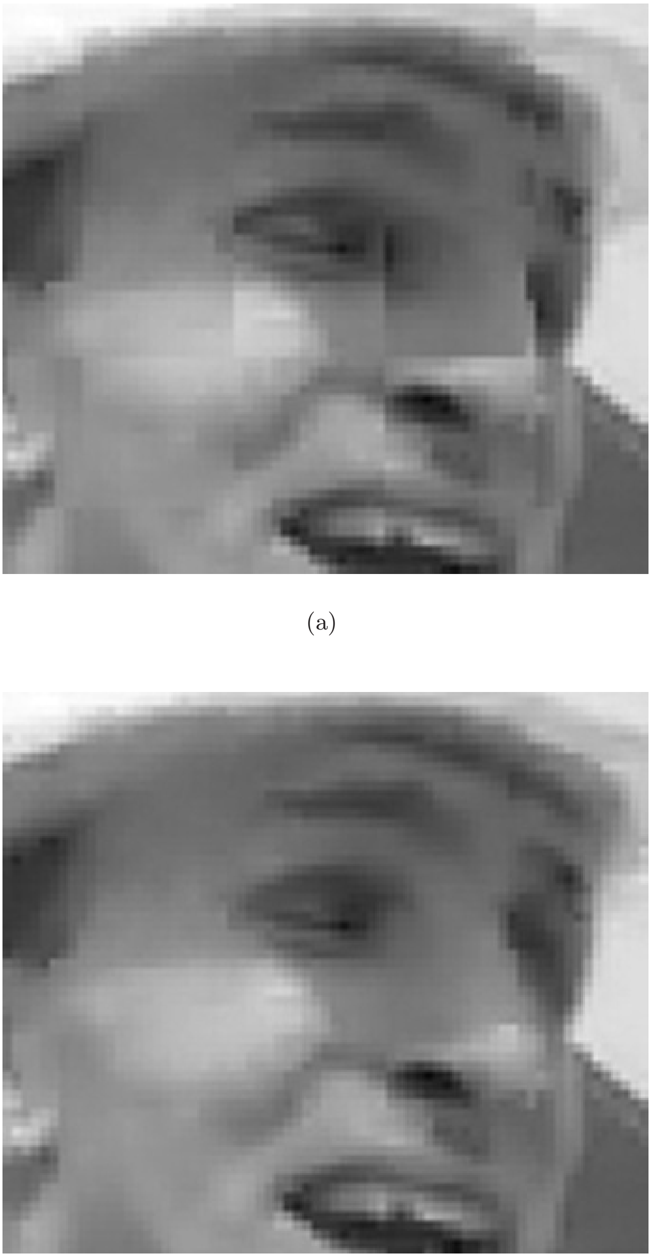
sequence	format	bitrate	w/o FMO			FMO-1		
			d=0	d=2	d=4	d=0	d=2	d=4
container	CIF	128k	31.00	30.98	30.95	30.59	30.50	30.26
paris	CIF	128k	23.98	23.96	23.64	26.10	26.47	26.22
foreman	CIF	128k	27.09	27.10	27.12	28.23	28.13	28.08
hall monitor	CIF	128k	30.10	29.98	30.08	31.64	31.55	31.42
akiyo	QCIF	64k	39.60	39.59	39.49	40.69	40.59	40.39
coastguard	QCIF	64k	25.49	25.51	25.45	27.77	27.76	27.64
foreman	QCIF	64k	29.41	29.39	29.40	31.18	31.15	31.06
hall monitor	QCIF	64k	33.96	33.97	33.85	34.87	34.61	34.19

5 Experimental Result

The proposed TEC scheme is evaluated under common test condition for packet-switched video service over 3G mobile transmission, which is contributed to the H.264 standardization group JVT (Joint Video Team) [5]. For simulating radio channel conditions, bit error patterns are provided by the common test condition [5]. The bit error patterns are captured between physical layer and the RLC/RLP layer over different real or emulated mobile radio channel. We limit the NAL unit size to 80 bytes because practically the maximum transfer unit (MTU) size in existing mobile network such as UMTS is less than 100 bytes. It turns out that the error rate should be increased as packet length is increased [4]. We also followed the bitrate and all other conditions as specified in [5]. All TEC schemes containing the proposed one are implemented in the H.264 reference codec, JM (Joint Model) version 7.3 [6]. To assist the error concealment technique, dispersed slice structure (FMO mode-1) is applied in this experiment.

Table 1 shows PSNR values of proposed method with regard to each depth of side region, d. It is obvious that the TEC performance under dispersed slice structure (FMO-1) is much higher than non FMO case. In case that d is zero, the performance of proposed scheme is identical to SMT which is conventional TEC scheme. As depicted in Table 1, PSNR value of proposed method is decreased gradually as d is increased. Because the side region overlapping affects the PSNR degradation in correctly decoded neighboring blocks of the lost block. However, it is not a big problem because the PSNR degradation is less than 0.1 dB on average and our purpose is to enhance the visual quality.

As shown in Fig. 4, we can see that the proposed one provides substantial improvement in subjective quality comparing conventional TEC scheme, SMT [3]. For the value of depth in extended block, d, 4 is used in this case. Comparison with the SMT scheme shows that the blocking artifact is removed effectively by overlapping side area in the proposed method especially in homogenous region. However the blocking artifact in fragmented blocks which come from the concealment with incorrect motion vector can not be removed.



(a)

(b)

Fig. 4. Comparison of subjective quality bewteen (a) SMT scheme and (b) the proposed one

6 Conclusion

In this paper, a new temporal error concealment scheme is proposed to compensate the discontinuity at lost macroblock boundary. Since the deblocking process can not be performed in concealed macroblock, the blocking artifact affects considerable subjective quality degradation. It is proved in the simulation result that our proposed scheme can compensate the discontinuity effectively by overlapping neighboring pixels of between lost block and reference block with weighting factor. It is also shown that FMO technique facilitates the performance of conventional error concealment including the proposed one.

References

1. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG: Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC). Doc. JVT-G050r1 (2003)
2. Lam, W., Reibman, A., Lin, B.: Recovery of lost or erroneously received motion vectors. Proc. ICASSP, vol. 5, (1993) 417–420
3. Zhang, J., Arnold, J., Frater, M., Pickering, M.: Video error concealment using decoder motion vector estimation. Proc. of IEEE TENCON - Speech and Image Technologies for Computing and Telecommunications. vol. 2, (1997) 777–780
4. Stockhammer, T., Hannuksela, M., Wiegand, T.: H.264/AVC in Wireless Environments. IEEE Trans. Circuits Syst. for Video Tech., vol. 13, No. 7, (2003) 657–673
5. Varsa, V., Karczewicz. M., Roth, G., Sjóberg, R., Stockhammer, T., Liebl, G.: Common test conditions for RTP/IP over 3GPP/3GPP2. ITU-T SG16/Q6 Doc. VCEG-N80, (2001)
6. http://bs.hhi.de/˜suehring/tml/download/jm73.zip

Spatio-temporally Adaptive Regularization for Enhancement of Motion Compensated Wavelet Coded Video[*]

Junghoon Jung, Hyunjong Ki, Seongwon Lee, Jeongho Shin, Jinyoung Kang,
and Joonki Paik

Image Processing and Intelligent Systems Lab.
Department of Image Engineering,
Graduate School of Advanced Imaging Science, Multimedia, and Film,
Chung-Ang University,
221 Huksuk-Dong, Tongjak-Ku, Seoul 156-756, Korea,
paikj@cau.ac.kr, http://ipis.cau.ac.kr

Abstract. The three-dimensional (3D) wavelet transform with motion compensation is a promising video coding algorithm with very high compression rate because of its spatial and temporal decorrelation. However, it still suffers from image degradation such as ringing artifats due to the loss of high frequency components by quantization. In this paper, we present an iterative regularized enhancement of the motion-compensated 3D wavelet coded video. The enhancement includes the adaptive implementation of the constraints for the regularization by selectively suppressing the noise along with the corresponding edge direction. The proposed algorithm efficiently reconstructs images defected by the three-dimensional wavelet transform.

1 Introduction

The growing demands on the quality of various video applications easily overwhelm current communication and storage technology. Due to the limited capacity of transmission bandwidth and storage devices which are available for most consumers, more advanced video compression methods than current standards such as MPEGs have been studied in academia and industry. Especially, the high quality video coding with low bit-rate is important for the video conferencing, videophone, etc. However, the bit rate often sacrifices the quality of image in video communication systems. Only a good encoder that removes the redundancy in both spatial and temporal correlations can meet this tough requirement. As an advanced coding scheme, the three-dimensional (3D) wavelet-based video coding that is an extension of spatial domain wavelet transform to the temporal domain has recently been proposed in [1,2,3,4,5].

[*] This work was supported in part Korean Ministry of Science and Technology under the National Research Lab. Project and in part by Korean Ministry of Education under Brain Korea 21 Project.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 232–240, 2004.

The coding efficiency for 2-D images of wavelet transform has already been shown in [6,7]. The wavelet-based coding for 2-D images has advantages over the conventional block DCT-based coding such as JPEG. First, wavelet transform processes the image in whole and consequently avoids the blocking artefact which is inevitable in block DCT-based coding. Second, the wavelet transformed data contain both frequency and spatial information. Thus, the wavelet transform achieves the higher energy compaction than other transforms.

In 3D video coding, wavelet transform is performed in the temporal domain as well as in the spatial domain. Namely, the input video signal is decomposed into spatio-temporal subbands by motion-compensated temporal filtering and the spatial-domain wavelet transform. The advantages of the 3D-wavelet decomposition with motion compensated temporal filtering include; (i) the temporal scalability can be achieved. The scalability refers to the methods which take some parts of the compressed bit-stream and decode the pictures at different quality levels [5]. (ii) The even higher energy compaction can be obtained.

The wavelet-compressed images, however, suffer from coding artifacts such as ringing artifacts at a higher compression rate, because quantization in high frequency subbands causes the loss of transform coefficient accuracy. Ringing artifacts appear as small ripples around the edge, and results in over- and undershoots in the lowpass filter response. In case of the temporal compression in 3D wavelet compressed video, if the intensity values of the same position on consecutive frames vary abruptly, ringing artifacts and motion blur appear at the same time. Ringing artifacts exist in the smooth areas of the frame. On the other hand motion blur occurs in dynamic regions and looks like the afterimage of an edge.

In this paper we propose a new algorithm that reduces such coding artifacts in decoded video by using adaptively iterative regularized restoration method based on the constraints of the 3D wavelet-based video coding. The proposed algorithm adaptively applies the different types of highpass filters to the regularized restoration according to edge directions.

This paper is organized as follows. Section 2 presents overview of the 3D wavelet compression video system. In Section 3 we propose the image restoration algorithm for 3D compressed video. Finally, Sections 4 and 5 discuss the experimental results and conclusions, respectively.

2 Three-Dimensional Wavelet-Based Video Compression System

The 3D wavelet transform with motion compensation [2,3,4] is basically a modifed version of typical three-dimensional (i.e., spatio-temporal) wavelet transform. The wavelet transform first decomposes two successive frames along the temporal axis, where the second frame is the motion compensated version. Then spatial wavelet coefficients are decomposed and quantized.

In the 3D wavelet-based video compression system, motion compensation reduces temporal redundancy by making two successive frames alike. More specifically, the temporal filtering in motion-compensated temporal analysis is performed along the motion trajectory [1,2]. The temporal analysis decomposes the frames with the two-tap Haar wavelet. The Haar wavelet is used because its short length is suitable for short signal decomposition. A longer tap temporal filter would result in a potential delay problem in the practical implementation of the system.

Once a video frame is made a pair with the motion-compensated consecutive second frame, the temporal wavelet filter decomposes them into the temporal low- and high-frequency subbands (L and H respectively). Then the L frames are grouped together and decomposed again into temporal LL and LH subbands at the second stage. Since the temporal decomposition transfers the energy of frames to the low frequency subbands, the lowest frequency subband contains most signal energy. At the lowest temporal level, we apply motion-compensated prediction to the t-LL subbands [1].

After the second level of temporal wavelet decomposition and motion prediction, the 3D wavelet decomposition results in intra t-LL, predictive t-LL, t-LH, and t-H. These frames are further spatially decomposed. As a result, the spatial correlations can be removed by applying a two-dimensional wavelet transform to the subbands resulting from motion compensated temporal transform subbands. The whole procedure is shown in Fig. 1.

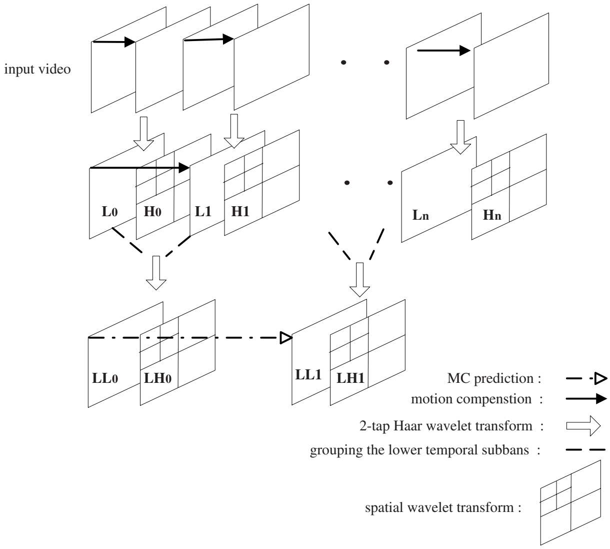

input video
L0
H0
L1
H1
Ln
Hn
LL0
LH0
LL1
LH1
MC prediction :
motion compenstion :
2-tap Haar wavelet transform :
grouping the lower temporal subbans :
spatial wavelet transform :

Fig. 1. Overview of the 3D wavelet video compression system

3 Enhancement Algorithm

We can define a model for the degradation system for the 3D wavelet-based video compression as

$$y = Dx \tag{1}$$

where, x represents the original video, y the compressed video, and D the degradation process due to the entire 3D wavelet-based coding process. The entire wavelet-compression and decompression system is depicted in Fig. 2.

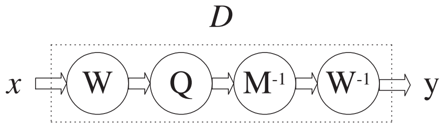


Fig. 2. The degradation model for 3D wavelet video compression system

According to Fig. 2, equation (1) can be rewritten as

$$y = Dx = W^{-1}M^{-1}QWx, \tag{2}$$

where W represents the 3D wavelet transform with motion compensation, W^{-1} its inverse transform. the quantization process, denoted by Q is performed in the encoder part and is further divided into two successive operations, division M and rounding R as

$$Q = RM. \tag{3}$$

The inverse quantization matrix M^{-1} that exists in the decoder simply represents the inverse of the division matrix M. The quantization operator, Q, can be realized in many different forms according to the quantization strategies for image compression methods [7]. The rounding operation, R in (3), is nonlinear and many-to-one mapping operator and plays a significant role in the entire degradation process because it is irreversible in the decoder.

The 3D wavelet compressed video suffers from ringing artifacts due to the loss of transform coefficient accuracy in higher subbands. Motion blur along the motion trajectory is another coding artifacts. In order to remove such coding artifacts, we solve equation (1) by the regularized method [8].

3.1 Formulation of Regularization

According to the regularization theory, we can obtain the regularized solution by minimizing the functional [8].

$$f(x) = ||y - Dx||^2 + \lambda||Cx||^2, \tag{4}$$

where D represents the degradation model of the 3D wavelet video compression. C represents the high-pass filter and λ the regularization parameter. The first term refers to the data compatibility. The second is the smoothness constraint term whose amount is controlled by λ.

For enhancing 3D wavelet-coded video, we must consider the highpass filter along the temporal axis as well as in the spatial domain. Thus we need additional constraint for equation (4), such as

$$f(x) = ||y - Dx||^2 + \lambda_1 ||C_s x||^2 + \lambda_2 ||C_t x||^2 \quad \text{subject to} \;\; x \in P, \tag{5}$$

where C_s represents the highpass filter applied to the spatial domain, each frame and C_t is performed along the temporal axis. P represents a hard constraint such as a clipping function that limits the intensity values of the recovered image within the range defined by a quantizer for the each iteration of the regularization process [13].

We can obtain the solution to minimize equation (5) by using the symmetric property of the degradation model [12], as

$$x^{k+1} = x^k + \beta\{y - (D + \lambda_1 C_s^\top C_s + \lambda_2 C_t^\top C_t)x^k\}. \tag{6}$$

In this solution, we assume that the degradation model is approximately linear.

3.2 Spatial Constraints

In the proposed regularization algorithm, we implement the spatially adaptive constraints for the directional highpass filter which is presented in [11,12]. The method uses a set of M different highpass filters, C^m, for $m = 1, \ldots, M$. Here, M is the number of the edge directions. It selectively suppresses the high frequency component along only the corresponding edge direction. The first regularization parameter λ_1 is adaptively produced by means of wavelet transform coefficients in high frequency bands [8]. We also implement adaptive highpass filters C^m, which use one of directional filters whose direction is determined by comparing the absolute value of transform coefficients. For example, each pixel in the image is classified as one of monotone, horizontal edge, vertical edge, and two diagonal edges. Then we develop the solution in the adaptive manner [12]. By extending equation (6) with directional information, we can have

$$\begin{gathered} x^{k+1} = x^k + \beta(y - \sum_{m=1}^{M} I_m T_m x^k) \\ \text{where} \;\; T_m = D + \lambda_1 {C_s^m}^\top C_s^m + \lambda_2 C_t^\top C_t, \end{gathered} \tag{7}$$

where I_m is a matrix whose diagonal components are equal to 1 or 0, and $\sum_{m=1}^{M} I_m = I$. This results in choosing the appropriate high-pass filter along the corresponding edge.

3.3 Temporal Constraints

In the degradation process, we can obtain the motion vectors along the temporal axis and consider it to the temporal constraint. Because the motion vectors are estimated between two successive frames, the temporal constraint, C_t, is applied to both frames and suppresses the high frequency along the motion vector. It is given as

$$C_t = [-0.25, 0.5, -0.25]^{\top}. \tag{8}$$

4 Experimental Results

Widely used test video sequences, Suzie and Foreman at QCIF resolution were used for experiments. The enhancement algorithm is applied to groups of 16 frames. Motion-compensated temporal analysis was performed by the two-tap Haar wavelet and the decomposition in the spatial domain used Daubechies 9/7 filters. We performed 2 levels of temporal analysis and 2 levels of spatial wavelet

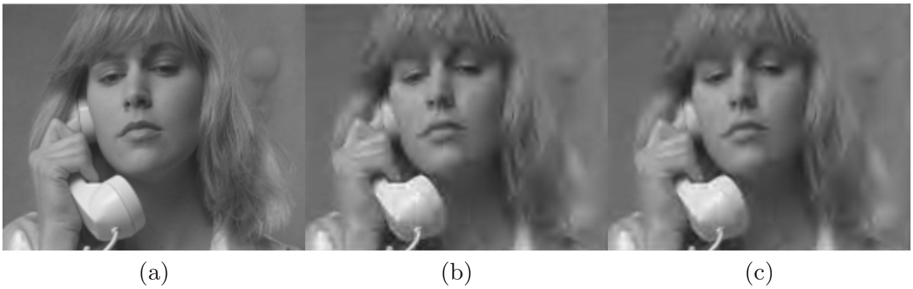


Fig. 3. (a) Original frame of the video (suzie), (b) the degraded frame by 3D wavelet video compression system and (c) the reconstructed frame by the proposed enhancement algorithm

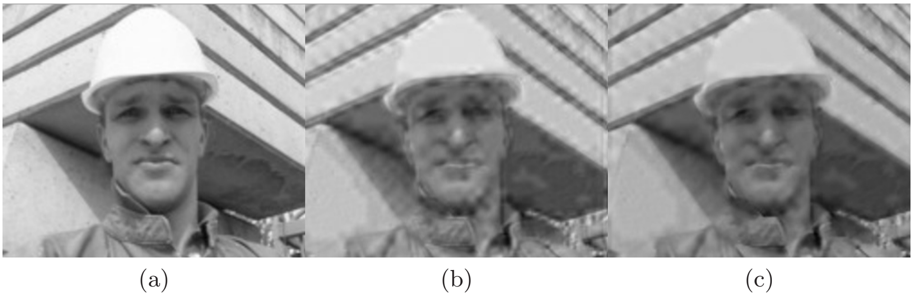
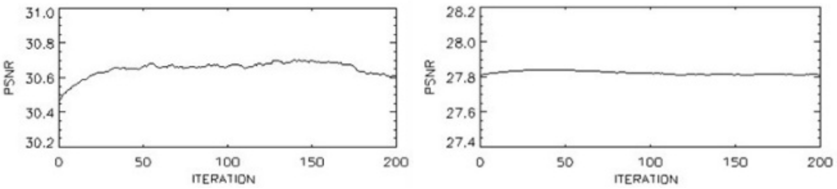
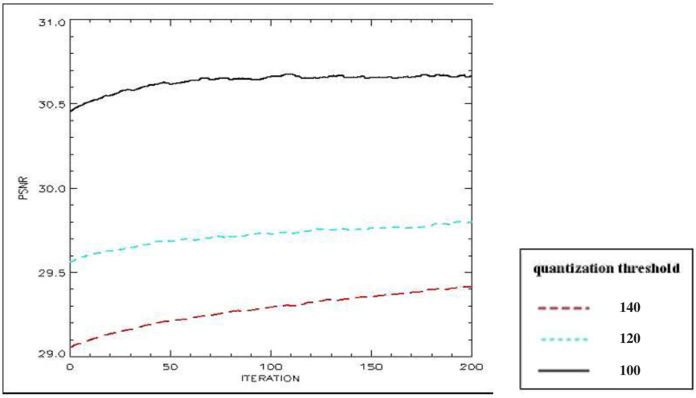


Fig. 4. (a) Original frame of the video (foreman), (b) the degraded frame by 3D wavelet video compression system and (c) the reconstructed frame by the proposed enhancement algorithm



(a) (b)

Fig. 5. PSNR vs. iteration: (a) Suzie and (b) Foreman



(a) (b)

Fig. 6. PSNR vs. iterations according to the quantization threshold (Suzie)

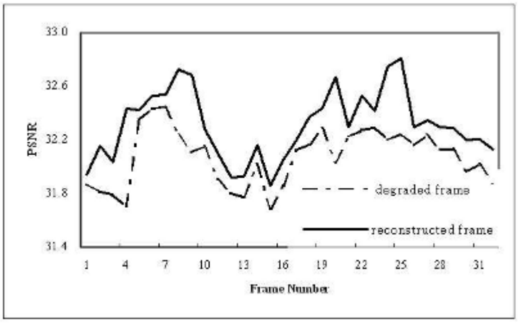


Fig. 7. PSNRs in two GOFs (Suzie)

decompositions. The parameters in the enhancement algorithm were determined experimentally, $\beta = 0.3$, $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ in case of Suzie image, $\beta = 0.2$, $\lambda_1 = 0.1$ and $\lambda_2 = 0.3$ in case of Foreman image.

A frame of the original Suzie sequence is selected and shown in Fig. 3(a) to demonstrate how the result image of our algorithm looks. Fig. 3(b) shows the same frame which is compressed and decompressed by the 3D wavelet-based coding with motion compensation. The decoded frame has the motion blur and ringing artifacts in region where the movement occurs. The enhanced frame with our proposed restoration algorithm is shown in Fig. 3(c). Compared with Fig. 3(b), the reconstructed frame has clearly less ringing artifacts and motion blur in the right hand and phone. In the same way, Figure Fig. 4(a), (b) and (c) show the experimental results of the Foreman sequence.

Quantization procedure is simply performed by cutting the wavelet transform coefficients below absolute magnitude. Fig. 5 shows PSNR vs. iterations when the wavelet transform coefficients below absolute magnitude 50 are quantized. How various quantization levels affect PSNR vs. iterations is seen in Fig. 6. The coefficients are quantized and compared at three different thresholds, 100, 120, and 140. The higher the threshold value is, the faster PSNR increases. Since higher threshold value implies the compacter final coded stream, the results confirm that we can effectively enhance the low bit-rate compressed images with our proposed algorithm. Finally, Fig. 7 shows the trend of PSNR on each frame from two GOF's with the Suzie sequence.

5 Conclusions and Future Works

Video coding at low bitrate condition inevitably accompanies coding artifacts in the decoded video sequence. Restoration algorithm is often used to improve the quality of decoded images. In this paper, we propose an adaptive regularization algorithm for enhancing 3D wavelet coded video with motion compensation. As shown in experimental results, the proposed restoration algorithm can efficiently restore the degraded image by reducing the coding artifacts while the details are preserved.

Because the compression method of the video is more complicated than the still image, we will consider additional constraints according to the degradation model in the future research. We also are going to modify the algorithm for real-time implementation. For example, the degradation model for a GOF can be decomposed and approximated into the individual degradation model for each frame. The enhancement procedure can be performed on each frame, then the enhancement of sequence can be implemented in real-time framework as presented in [11].

References

1. Choi, S.J., Woods, J.W.: Motion-Compensated 3-D Subband Coding of Video. IEEE Ttrans. on Image Processing **8**(2) (1999) 155–167
2. Ohm, J.R.: Three Dimensional Subband Coding with Motion Compensation. IEEE Trans. on Image Processing **3**(5) (1994) 559–571
3. Podilchuk, C.I., Jayant, N.S., Farvardin, N.: Three-Dimensional Subband Coding of Video. IEEE Trans. Image Processing **4**(2) (1995)
4. Levy, I.K., Wilson, R.,: Three Dimensional Wavelet Transform Video Compression. Proc. of the IEEE Multimedia Systems **2** (1999) 924–928

5. Pesquet-Popescu, B., Benetiere, M., Bottreau, V.: Embedded Color Coding For Scalable 3D Wavelet Video Compression. Proc. SPIE Visual Commun. Image Processing (2000)
6. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image Coding Using Wavelet Transform. IEEE Trans. Image Processing **1** (1992) 205–220
7. Shapiro J.M.: Embedded Image Coding Using Zerotrees of Wavelet Coefficients. IEEE Trans. Signal Processing **41**(12) (1993) 3445–3462
8. Katsaggelos, A.K.: Iterative Image Restoration Algorithms. Optical Engineering **28**(7) (1989) 735–748
9. Jain, A.K.: Fundamentals of Digital Image Processing, Prentice-Hall (1989)
10. Mallat, S., Zhong, S.: Chracterization Of Signals From Multiscale Edges. IEEE Trans. Pattern Analysis and Machine Intelligence **14**(7) (1992)
11. Jung, J.H., Shin, J.H., Paik, J.K.: Spatio-temporally Adaptive Image Sequence Interpolation. Proc. 1998 Int. Tech. Conf. Circuits, Systems, Computers, Communications **1** (1998) 43–46
12. Jung, J.H., Joung, S.C., Paik, J.K.: Regularized Constrained Restoration of Wavelet Compressed Image. Proc, SPIE Visual Commn. Image Processing (2000)
13. Yang, Y., Galantsanos, N.P., Katsaggelos, A.K.: Regularized Reconstruction to Reduce Blocking Artifacts of Block Discrete Cosine Transform Compressed Images. IEEE Trans. Circuits Syst. Video Technol **3**(6), no. 6, pp. 421–432, 1993

ROI and FOI Algorithms for Wavelet-Based Video Compression

Chaoqiang Liu[1,2], Tao Xia[2], and Hui Li[1,2]

[1] Temasek Laboratories, Approximation and Information Processing,
National University of Singapore
[2] Centre for Wavelets, Approximation and Information Processing,
National University of Singapore

Abstract. Many techniques make great contributions to video compression with removal of spatial and temporal redundancy in and between frames. However, compressed video is still rather large for applications such as surveillance system. In order to compress more, in video compression techniques, region-of-interest (ROI) and frames-of-interest (FOI) codings could be addressed to set high priority to ROI or FOI by allocating more bits than others. Normally, locations of related coefficients for the reconstruction of the ROI are calculated according to the filter lenghth. However, it is not efficient. In this paper a novel wavelet-based ROI and FOI scheme is proposed. Simulation results show excellent performance.

1 Introduction

Recently video compression techniques have drawn much attention. Especially in such applications as video surveillance, video browsing, and so on. While conventional video compression techniques reduce redundancy of video frames, some other subtle approaches are introduced to improve compression ratio and coding efficiency further. Among them region of interest (ROI) technique is one of the most successful ones. ROI achieves higher subjective quality by spending most bits on the important/interested region and sacrificing the quality of less important part. Many schemes have been proposed [1]–[4], most of which use some common shapes or fixed shapes for ROI.

ROI technique is simple by predefining a mask of interested region. However, for transform based coding, the mask of ROI in transformed domain should be determined precisely. With the popularity of wavelet technology which is adopted by JPEG2000 as the suggested transform, the mask generation problem for compression based on wavelet transform is studied. New algorithms of ROI and FOI to suit video compression applications are developed. Once arbitrarily shaped ROIs are defined by user, generation of the ROI mask in wavelet domain is performed. The proposed ROI mask generation algorithm supports arbitrary shape ROIs and arbitrary combination of different wavelet filters. It can be used in applications more than compression, such as other applications using translation invariant wavelet transform. Meanwhile, FOI technique is another

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 241–248, 2004.

useful approach for some applications, such as surveillance application, as people have no interest in the portion without interested contents. The FOI algorithm is proposed as well in this paper for video compression.

2 ROI/FOI Algorithms for Wavelet Based Video Compression

The functionality of ROI is important in applications where certain parts of the image are of higher importance than others [5]–[7]. In such a case, these regions need to be encoded at higher quality than the background. During the transmission of the image, these regions need to be transmitted first or at a higher priority. In the general ROI coding methods, after wavelet transformation, the coefficients associated with the region-of-interest will be transferred ahead of those associated with the background. Therefore, when an image is coded with an emphasis of ROI, it is necessary to identify the wavelet coefficients required for the reconstruction of the ROI. Thus, the ROI mask is introduced to indicate which wavelet coefficients have to be transmitted exactly in order for the receiver to reconstruct the ROI. A lot of research works have been studied about the ROI mask. Some are in normal or fixed shape, such as the rectangle. Others are for arbitrary shape but only for some special transforms [8]. Usually, in the general scaling based method, the wavelet transform is applied to the image at the encoder and the resulting coefficients not associated with the ROI are scaled down (shifted down) so that the ROI-associated bits are placed in higher bit planes [9].

The mask in wavelet domain is a map pointing out all the related coefficients for the reconstruction of the ROI. Normally, to find the mask in different scales, the inverse wavelet transformation is studied [10]. The corresponding locations of the coefficients in next scale are calculated from the current scale. For instance, For the 5/3 filter, it can be seen that to reconstruct $X(2n)$ and $X(2n+1)$ losslessly, coefficients $L(n), L(n+1), H(n-1), H(n), H(n+1)$ are needed. Hence

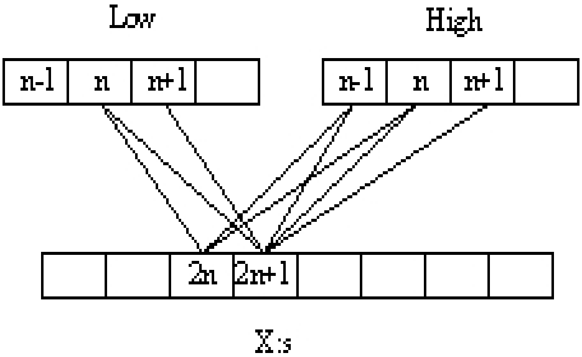


Fig. 1. The inverse 5/3 filter

if $X(2n)$ or $X(2n+1)$ are in the ROI, the listed low and high subband coefficients are in the mask, as shown in Figure 1.

However, the above method has some obvious limitations. For instance, the 2-D masks in different scales are not easy to compute when there are several regions of interest. While there are several different filters involved or prefiltering operations are needed for multiwavelet compression [11], the existing algorithms will not work. Therefore, a direct and automatic approach for ROI mask generation is suggested. From a ROI mask, a set of functions is computed for different scales and different subband components by imposing similar forward wavelet transform. These functions indicate the masks of ROI in wavelet domain automatically with minor futher operations. The process for a 2-level decomposition is shown in Figure 2.

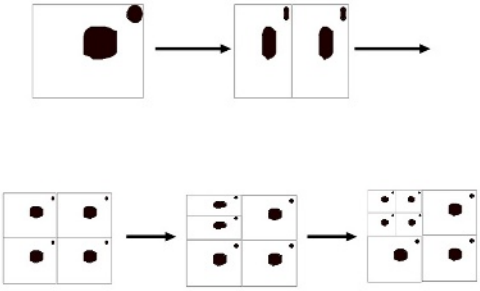


Fig. 2. ROI mask generation in 2-level DWT domain

2.1 Mask Generation for ROI

Besides the scheme of coding, the main effort for ROI is how to generate the mask. To understand this problem better, it is stated in a more general case.

Let us assume to impose a wavelet operation on R^n, if the region $\Omega \subset R^n$ is of interest, now the question lies in how to find the corresponding regions in wavelet domain.

To solve this problem, first, the characteristic function $\chi_\Omega(x)$ is defined as usual,

$$\chi_\Omega(x) = \begin{cases} 1, & \text{if } x \in \Omega \\ 0, & \text{if else} \end{cases} \tag{1}$$

With the aid of a set of functions $\{g_i(x)\}_{i\in\Lambda}$ defined below the mask generation is straightforward,

$$g_i(x) = (\widetilde{W_i} \circ \chi_\Omega)(x) + \widetilde{I}_i \chi_\Omega(x) \quad i \in \Lambda \tag{2}$$

where $\widetilde{W_i}$ stands for the wavelet operator for ith subband and Λ is the index set of all subbands, here subbands also include the subbands in different scales.

It is obvious that $\overline{\Omega}_{m,i}$ is the corresponding region of Ω in wavelet domain, where $\Omega_{m,i} = \{x \mid g_i(x) > 0\}$ and $\overline{\Omega}_{m,i}$ is the closure of $\Omega_{m,i}$.

Remarks

1. Wavelet operator $\widetilde{W_i}$ in (2) is generalized instead of a real operator. It might be a wavelet transform equipped with downsampling operation, while $\widetilde{I_i}$ is identity operator equipped with downsampling operation respectively.
2. The closure operation for Ω_m can be implemented with some other techniques in real applications.
3. When $n = 2$, Ω is the region of interest in the image/video processing/coding application. Accordingly, $\overline{\Omega}_{m,i}$ is the corresponding ROI mask in wavelet domain in ith subband.

Therefore algorithm for the mask generation is as follows.

Algorithm 1

1. From the region of interest Ω construct characteristic function $\chi_\Omega(x)$,
2. For all $i \in \Lambda$, applying wavelet operation $\widetilde{W_i}$ to obtain $\widetilde{W_i} \circ \chi_\Omega$, then compute the function $g_i(x)$,
3. Computing $\Omega_{m,i}$, then get the closure of $\Omega_{m,i}$ as $\overline{\Omega}_{m,i}$.

Remarks

1. This algorithm can accompany with the normal wavelet transform denoted by W_i. The only difference here is that the filters used in $\widetilde{W_i}$ are the dual filters of those used in W_i (in other words, inverse filters are used).
2. The closure operation in step 3 differs for different cases, in image and video case, the morphology operators could be used to fill the hole in $\Omega_{m,i}$ to get $\overline{\Omega}_m$.

The algorithm 1 gives automatic generation of the ROI mask in wavelet domain. The advantages include its ability to generate the mask for any shape, any wavelet operation or combination of operations with different wavelet filters and adaptability with the standard transform operation.

2.2 ROI Video Compression

As mentioned above, ROI mask is formed automatically. The algorithm is applicable to different filters and requirements, and can be combined with existing useful techniques, such as scaling down (shifting down) and MAXISHIFT [13]–[16]. Based on the aforementioned mask generation algorithm, the ROI image/video compression algorithm is completed as below.

Algorithm 2

1. Get the ROI map.
2. Apply wavelet transform to image,
3. Apply similar wavelet transform (using inverse filters) on the ROI map to get mask using algorithm 1
4. Coding the coefficients in wavelet domain with scaling technique to emphasis the information for the ROI.
5. Quantization and coding the wavelet coefficients.

Based on the observation that the ROI map in image domain for video compression has huge temporal redundancy, the ROI mask generation algorithm for video compression could be further improved. As the mask generation procedure is a linear operation, it is only necessary to compute the difference of the masks of two ROI maps in the successive frames. Even though the ROI is detected dynamically by some other techniques [17][18], this property can be exploited of ROI masks to generate ROI masks for video compression more efficiently. Moreover, if the moving objects in a video clip can be detected and marked as ROI portions, such a ROI video will improve the visual perception quite a lot due to the less sensitivity of human eyes to the still parts in a video.

2.3 FOI Video Compression

Besides, Frame-of-interest(FOI) can also contribute to video compression with the intelligent motion estimation algorithm which can detect the interested frames. Then FOI allows better quality during certain time interval of compressed video, while permits low quality of other periods of the video. The following Figure 3 explains the idea of FOI. There is a high quality compressed video. After FOI procedure, it forms a FOI video sequence, in which certain time interval still keeps high quality and same display period as in the original

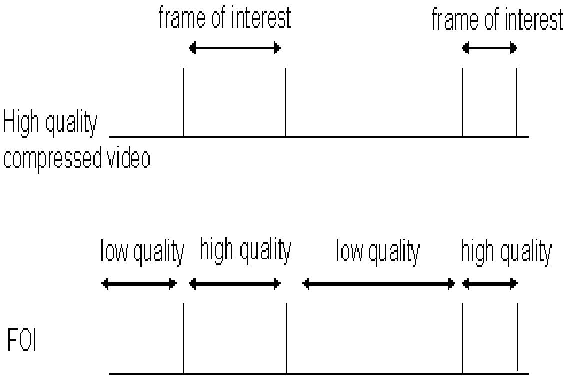


Fig. 3. FOI video compression

video, while other time intervals may be low quality and even shorter display period compared to that in the original video, which means that the video abstract technique may also be applied to that period to reduce the video display time.

An application of FOI video compression is developed, in which specified intervals of the video can be displayed with higher quality than others after compression. Frames of interest could be freely specified by users. Moreover, the ROI video compression technique in above section could be combined with FOI scheme. Imagine that the ROI algorithm is used in those low quality video parts, the visual quality of the whole video is improved undoubtable. Or furthermore, some video abstract technique [19] can be applied into those low quality time intervals, to save the temporal space by leaving only few key frames.

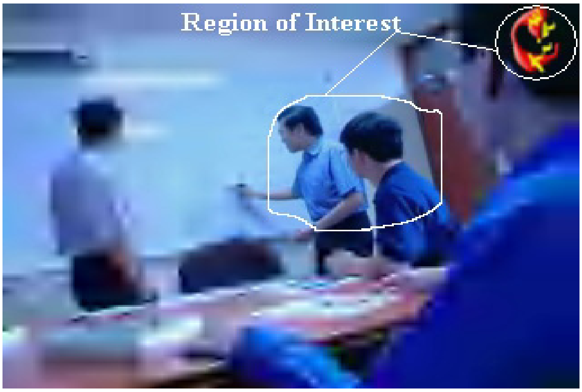


Fig. 4. ROI in video

3 Experiments and Simulation Results

The simulation result is shown in Figure 4. The ROI mask is highlighted in the figure, which is in the same shape as the mask described in ROI mask generation, as shown in Figure. 2. It is clear that this ROI is of high quality all over the video. From such a point of view, uses could focus on certain region of the video that is of interest, for different kinds of filters, with or without downsampling. Meanwhile, the mask generation is quite simple, efficient and direct with only applying a certain type of wavelet transform on the ROI map.

4 Conclusions

In this paper, a new ROI/FOI scheme for video compression is developed. In this method, wavelet transform is directly applied to the ROI map to generate mask in each scale. The algorithm works not only for arbitrary shaped ROI but also for any combination of different wavelet filters and translation variant and translation invariant wavelet transform. It is also practical to different kinds of filters, with or without downsampling, and can be combined with many useful techniques. FOI mechanism is also introduced for the video compression, together with ROI. The video compression can be implemented for the cases of with some period of clips of low quality and some periods of clips of high quality only at the regions of interest. Together with other techniques such as the automatic target detection or moving object detection, video compression with much higher compression ratio and better human visual perception will be achieved. Simulation results are presented with excellent performance.

Acknowledgments. The authors would like to thank Defence Science and Technology Agency (DSTA) and Temasek Laboratories for support of the programme "Wavelets and Information Processing".

References

1. C.W. Lin, Y.J. Chang and Y.C.Chen, "Low-Complexity Face-Assisted Video Coding",Proc. ICIP2000, vol. 2, pp. 207–210, Sep. 2000
2. M. Morimoto,K. Matsumura and K. Fujii, "A hierarchical video compression method using object coding", World Automation Congress, 2002. Proceedings of the 5th Biannual, vol. 13 , pp. 345–350, June 2002
3. A. Eleftheriadis and A. J.Jacquin, "Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rate", Signal Processing: Image Communication, vol. 7, no. 4-6, pp. 231–248, Nov. 1995
4. K. Fung, N.Law, and W.Siu, "Region-based object tracking for multipoint video conferencing using wavelet transform", Consumer electronics, 2001 ICCE International Conference, pp. 268–269, June. 2001
5. J. Strom and P.C.Cosman, "Medical image compression with lossless regions of interest", Signal Processing, vol. 59, no. 2, pp. 155–171, June. 1997
6. J. Askelof, C. Christopoulos, M. Larsson Carlander, and F. Oijer, "Wireless image application and next-generation imaging", *Ericsson Review*, no. 2, pp. 54–61, 2001. Available http://www. ericsson.com/review/2001_02/
7. D. Santa-Cruz, T. Ebrahimi, M. Larsson, J. Askelof, and C. Christopoulos,"Region of interest coding in JPEG2000 for interactive client/server applications," in *Proc. IEEE* 3^{rd} *Workshop on Multimedia Signal Processing*, pp. 389–394, Sept. 1999
8. D. Nister and C. Christopoulos, "Lossless Region of Interest with Embedded Wavelet Image Coding", *Signal Processing*, vol. 78, no. 1, pp. 1–17, 1999
9. Z. Wang, and A.C. Bovik, "Bitplane-by-Bitplane Shift (BbBShift)—A Suggestion for JPEG2000 Region of Interest Image Coding", IEEE Signal Processing Letters, vol. 9, no. 5, May. 2002
10. C. Christopouls, A.N. Skodras, and T. Ebrahimi, "JPEG 2000 still image coding system: An overview", *IEEE Trans. Consumer Electronics*, vol. 46, pp. 1103–1127, Nov. 2000
11. T. Xia and Q. Jiang, "Optimal multifilter banks: design, related symmetric-extension and application to image compression", IEEE Transactions on Signal Processing, vol. 47, no. 7, pp. 1878–1889, July. 1999

12. J.M. Shapiro, “Embedded image coding using zerotree of wavelet coefficients”,IEEE Trans. Signal Processing, vol. 41, pp. 3445–3462, Dec. 1993
13. E. Atsumi and N. Farvardin, “Lossy/lossless region-of-interest image coding based on set partitioning in hierarchical trees”, Proc. IEEE Int. Conf. Image Processing, Chicago, IL, pp. 87–91, Oct. 1998
14. C.A. Christopoulos,J. Askelof, and M. Larsson, “Efficient methods for encoding regions of interest in the upcoming JPEG2000 still image coding standard”, IEEE Signal Processing Lett., vol. 7, pp. 247–249, Sep.2000
15. C.A. Christopoulos,J. Askelof, and M. Larsson, “Efficient encoding and reconstruction of regions of interest in JPEG 2000”, in Proc. X European Signal Processing Conf., Tampere, Finland, pp. 1133–1136, Sep.2000
16. C.A. Christopoulos,J. Askelof, and M. Larsson, “Efficient region of interest encoding techniques in the upcoming JPEG 2000 still image coding stadard”, in Proc. IEEE Int. Conf. Image Processing, vol. II, Vancouver, Canada, pp. 41–44, Sep.2000
17. D. Li,“Moving objects detection by block comparison”, Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conference, vol. 1 , pp. 341–344, Dec. 2000
18. R. Zaibi, A.E. Cetin,and Y. Yardimci, “Small moving object detection in video sequences”, Acoustics, Speech, and Signal Processing, 2000. ICASSP ’00. Proceedings. 2000 IEEE International Conference, vol. 6, pp. 2071–2074, June. 2000
19. C. Liu, T. Xia, and H. Li, “Time Oriented Video Summarization”, IEEE Trans. On Circuits and Systems for Video Technology, under review

Adaptive Distributed Source Coding for Multi-view Images

Mehrdad Panahpour Tehrani[1], Michael Droese[2], Toshiaki Fujii[3], and
Masayuki Tanimoto[3]

[1] Department of Information Electronics
Geaduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
mehrdad@tanimoto.nuee.nagoya-u.ac.jp
http://www.tanimoto.nuee.nagoya-u.ac.jp/~tanilab/mehrdad/index.htm
[2] Department of Information Electronics
Geaduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
droese@ieee.org
[3] Department of Information Electronics
Geaduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
{fujii,tanimoto}@nuee.nagoya-u.ac.jp

Abstract. We show that distributed source coding of multi-view images using adaptive modules-operation can come close to the Slepian-Wolf bound. In the systematic scenario considered, two parent nodes (PN1 and PN2) on sides and child nodes (CNs) are located between PNs, which are statistically depended. A PN sends the whole image whereas a CN only partially, using an adaptive distributed source coding at a rate close to H(CN|PN1, PN2). The proposed scheme allows independent encoding and jointly decoding of views. Experimental results show performance close to the information-theoretic limit. Furthermore, good performance of the proposed architecture with adaptive scheme shows significant improvement over previous work.

1 Introduction

Multi-view images of a scene can be used for several applications ranging from free viewpoint television (FTV) [1] to surveillance. Due to the enormous size of multi-view images, coding is one of the challenges to build such applications. Multi-view images are usually highly correlated in spatial domain and therefore spatial redundancy can be removed by encoding the information “differentially” with respect to an appropriate “reference”. The disadvantage of this method is an extra overhead on communication between those nodes. In a scenario with limited processing and communication abilities this method is not preferable. Work by Slepian and Wolf [2] shows that even if the sources are encoded independently, they can be fully reconstructed under certain conditions. In other

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 249–256, 2004.

words, the Slepian- Wolf theorem suggest that it is possible to encode statistically dependent signals in a distributed manner to the same rate as with a system where the signals are jointly encoded. Therefore, distributed source coding of multi-view images is preferable if there is a major constraint on individual camera node performance (i.e., energy, which is consumed by sensing and communication operations). However, approaching the Slepian-Wolf bound is still a challenging issue.

Some work has been carried out [3–7] in designing a distributed source coding but the performance is not close to information theoretic bound. Aaron et al [8] proposed compression with side information using turbo codes. This method approaches the theoretic bound, however it resembles our method in a different way.

We propose an adaptive distributed source coding method without inter-node communication for multi-view images; similar to the distributed source coding method proposed in [9,10] based on module operation. To perform the decoding task, disparity estimation is employed to compensate the scene geometry to provide the side information. Experimental results show performance close to the limit of information theory. Furthermore, the proposed architecture with adaptive scheme shows significant improvement over previous work.

The remainder of the paper is organized as follows: Section 2 describes the coding scheme of CNs in detail. Section 3 shows the experimental results. Finally, Section 4 is conclusions of this research.

2 Coding Method

The cameras in the multi-view system are grouped into correlated clusters. Each cluster of nodes is coded independently. The cluster size corresponds to the maximum allowable disparity, respectively the maximum distance of a camera pair. Figure 1 shows the individual coding of multi-view images. It demonstrates coding with two PNs and therefore is called PCP (PN, $\cdots$, CN, $\cdots$, PN). The arrows in Figure 1 show the view to be used to decode a CN at the joint decoder. CNs in PCP are decoded using virtual PNs (vPN), which are generated by the two PNs at the outer border of a cluster (i.e., geometry compensation).

Before describing the encoding/decoding algorithms it is essential to note the variables used throughout the following sections. "$n \times m$" describes a block to be encoded. "D" stands for the maximum gray level bound that is imposed on the multi-view image coding at each block. "Maximum disparity" stands for the

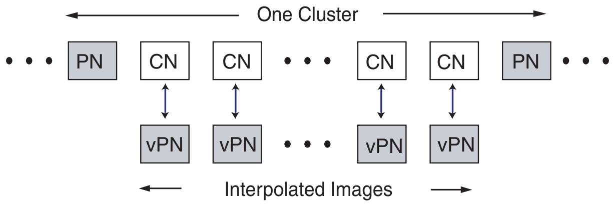


Fig. 1. Distributed source coding architecture, PCP

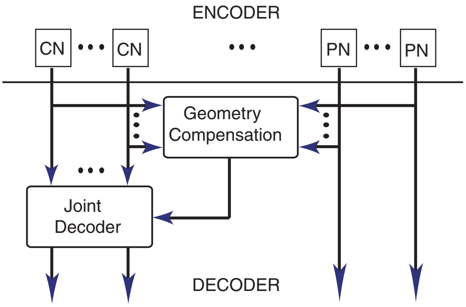


Fig. 2. Distributed source coder in general

number of pixels required to find all correspondences in a stereo setup. It also defines the size of a cluster. Figure 2 shows a block diagram of the proposed distributed source coder. Although CNs and PNs are set apart, in practice they are arranged one after another as shown in Figure 1. Due to no inter-node communication amongst cameras, the CN views are encoded independently at each node. The encoded data is transmitted to the joint decoder. At the joint decoder the side information from PNs is provided by the scene geometry, which is obtained by an area-based matching method of [11,12].

2.1 Encoding

The encoding of a CN works as follows: Pixels of each block are encoded with a "D" value. The adaptive value of the "D" is decided by using the average absolute gradients of a block in vertical and horizontal directions. It corresponds to the spatial frequency of the scene.

In our adaptive coding scheme, the higher spatial frequency, the higher "D" value is used. Based on the range, where the measured average gradient is, the adaptive "D" to encode a block is obtained. Table 1 shows the look up table to decide the adaptive "D" at each range. After choosing the adaptive "D" value, the pixels in a block are further encoded by applying a module-operation based on "D" to the pixel values. The same algorithm is applied to other blocks in the image. The encoded image using the adaptive scheme is called "coset image". The quality of the decoded image can be control by changing the average "D" value used for an image at encoder side. Multiplying a linear weighting factor (i.e., ≥ 0) to the measured average gradient does the controlling procedure. Note that the PN images are fully transmitted without any coding. The encoding flowchart is shown in Figure 3($left$).

Table 1. Look up table for adaptive distributed source coding

Gradient Range	0	1	2∼3	4∼7	8∼15	16∼31	32∼63	64∼127	≥128
D	1	2	4	8	16	32	64	128	256

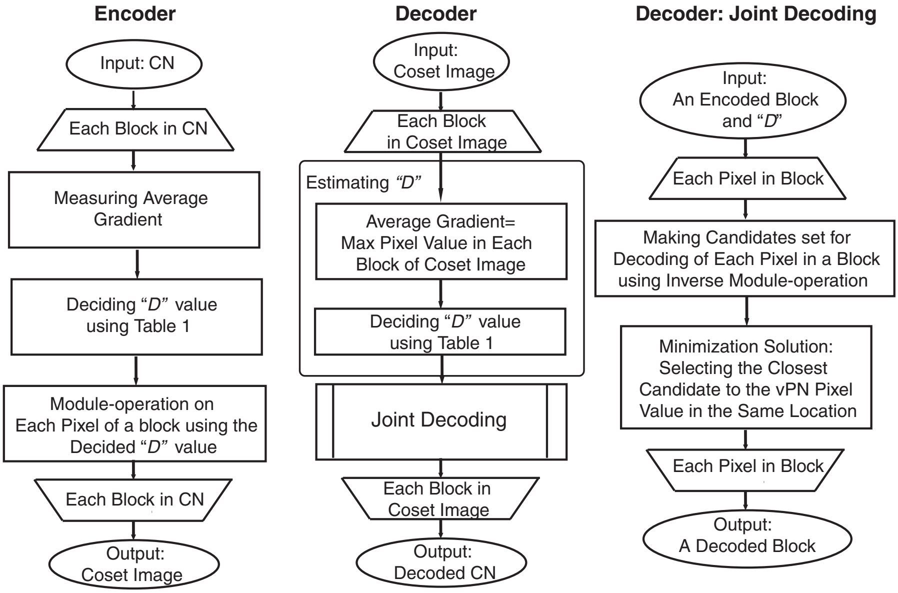

Encoder
Input: CN
Each Block in CN
Measuring Average Gradient
Deciding "D" value using Table 1
Module-operation on Each Pixel of a block using the Decided "D" value
Each Block in CN
Output: Coset Image
Decoder
Input: Coset Image
Each Block in Coset Image
Estimating "D"
Average Gradient= Max Pixel Value in Each Block of Coset Image
Deciding "D" value using Table 1
Joint Decoding
Each Block in Coset Image
Output: Decoded CN
Decoder: Joint Decoding
Input: An Encoded Block and "D"
Each Pixel in Block
Making Candidates set for Decoding of Each Pixel in a Block using Inverse Module-operation
Minimization Solution: Selecting the Closest Candidate to the vPN Pixel Value in the Same Location
Each Pixel in Block
Output: A Decoded Block

Fig. 3. Flowchart of (*left*)Ecoder and (*middle*)Decoder (i.e., (*right*)Joint Decoding) of Distributed source coding for PCP architecture

2.2 Decoding

After receiving the encoded data of CNs and full information of PNs at the joint decoder, the coset images of CNs should be reconstructed. Due to the module-operation applied to the coset image, decoding is necessary. This is done by using side information from the estimated scene geometry.

In PCP, a vPN image is generated [11,12] in the location of the CN as shown in Figure 1. The CN image is decoded by using the vPN's pixels at the same location as the coset pixels. The values of the coset image are decoded by applying an "inverse" module-operation, which is not unique. Therefore, that solution is chosen which minimizes the distance to the corresponding pixel of the other image (i.e., vPN).

However, decoding of the coset image is not possible, if the "D" value for each block is not known. To solve this problem, there are three ways as follow:

1. Sending "*D*" value for each block from encoder to decoder.

2. Estimating "*D*" value by using vPN image (i.e., side information). Table 1 is used for vPN image at decoder to estimate the "*D*".

3. Estimating "*D*" value by using CN image (i.e., coset image). The maximum coset value of each block refers to the range and then the "*D*" is decided using Table 1.

The first method due to overhead on the transmission rate is not preferable. Therefore, we would like to estimate the "D" value of each block to decode the coset image. Experimental results on different block sizes and image scenes show that the third method performance is nearly the same as the first method. Hence, we proposed to use the third method for decoding. The flowcharts of decoder and the joint decoding algorithms are shown in Figure 3($left$ and $right$), respectively.

3 Experiment

In this experiment, the data set consists of 3 views with 320×240 pixels per view of Cube&Doll, and 3 views with 384×288 per view of Tsukuba. In Cube&Doll the camera interval is 15mm and the distance from camera to object measures about 30cm. In Tsukuba the maximum disparity between the two furthest views is 60 pixels. All experiments are carried out on luminance component only. The reconstruction performance and information-theoretic bound of the adaptive coder for both test data sets are compared with a coder which is using the same "D" value for all pixels in CN image (i.e., fixed coder [10]). Reconstruction quality of a CN is measured in term of peak-signal-to-noise-ratio (PSNR) and compared with vPN quality.

To build up PCP architecture, two PNs are set apart with maximum disparity of 60 pixels and the middle view is CN. This is equal to a stereo setup with a camera baseline of 30mm. The block size in adaptive coder is 4×4. Due to geometry compensation at the joint decoder (i.e., vPN image), the proposed distributed source coder should perform better than vPN. Therefore in Figure 4

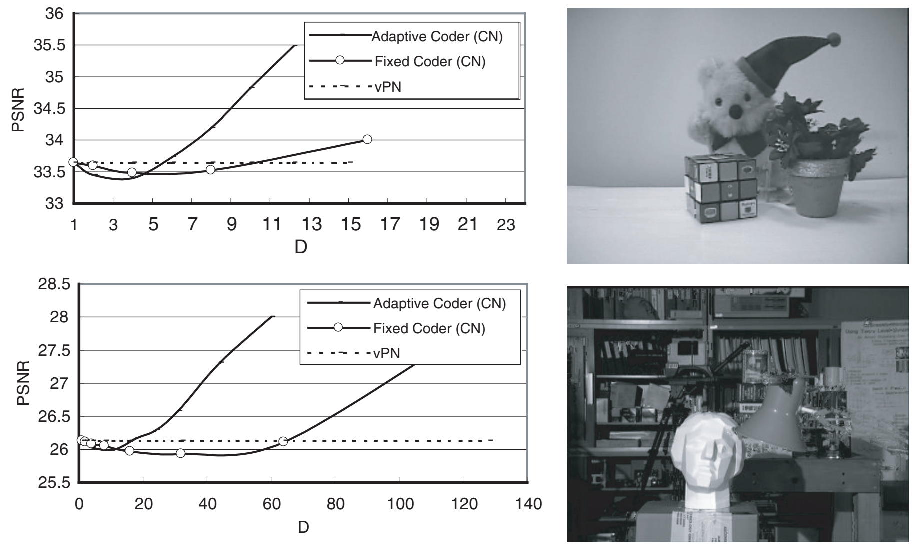


Fig. 4. PSNR vs. "D" for adaptive and fixed coders in comparison with vPN (*up*) Cube&Doll, 36.37dB for average D=16 (*down*) Tsukuba, 27.51 for average D=48

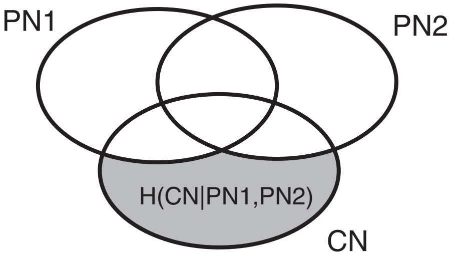


Fig. 5. Statistical dependence of PN1, PN2, and CN with entropy

the decoded CN quality has compared with vPN quality. Figure 4 shows PSNR curves vs. "*D*" for adaptive and fixed coders in comparison with vPN, and examples of reconstructed image of CN after decoding. In adaptive scheme, the "*D*" value shows the average "*D*" value. The proposed adaptive coder has gain over vPN quality with lower value of average "*D*" in comparison with fixed coder. Furthermore, the experimental result for Tsukuba shows a significant improvement due to the complexity of the scene.

According to the aforementioned importance of adaptive distributed source coding of multi-view images, the information-theoretic bound proposed by Slepain and Wolf is appealing. In this part of experiment, the CN rate is compared with the rate proposed by Slepain and Wolf. The PN1 and PN2 are to the images located at outer border of a cluster in PCP architecture and the CN is the to be encoded image in the middle. Due to the coding architecture, H(PN1, PN2, CN) - joint entropy of PN1, PN2, and CN - and H(PN1, PN2) - joint entropy of PN1 and PN2 - should be measured. After measuring these values, H(CN|PN1, PN2)= H(PN1, PN2, CN)-H(PN1, PN2) can be calculated as shown in Figure 5.

The H(CN|PN1, PN2) value is the point at which the rate of CN (Rx) should be close after it is encoded (i.e., coset image -). Table 2 and Table 3 show the results for two different rates using Cube&Doll and Tsukuba, respectively. The tables shows the ratio of Rx and the ideal rate H(CN|PN1, PN2) with R1, and the ratio of the combined rate Rx+H(PN1, PN2) and the ideal combined rate H(PN1, PN2, CN) with R2.

The results mentioned in the tables show that by using the adaptive coding scheme the encoded rate can be closer to Slepain-Wolf bound than fixed coder. In Cube&Doll image when the average "D=22" the Rx is completely satisfied the theoretical rate, with 4.5dB gain over the vPN. Obviously, the Tsukuba has more complicated scene than Cube&Doll, therefore the rate inefficiency is higher than Cube&Doll. However, according to Figure 4 in Tsukuba the minimum required average "*D*" value for efficient decoding with gain over vPN has decreased much more than that of Cube&Doll.

Considering the performance of decoded CN shown in Figure 4 and the achieved rates of CN in Table 2 and 3 show that by applying the adaptive distributed source coding, we could improve the quality of the decoded result

Table 2. Rate Rx achieved by the adaptive scheme as compared to Slepian-Wolf bound (Cube&Doll)

	D=16		D=22		D=32		D=48	
	R1*	R2**	R1	R2	R1	R2	R1	R2
Fixed Coder	1.54	1.08	1.62	1.11	1.63	1.12	2.02	1.18
Adaptive Coder	0.85	0.97	0.98	0.99	1.26	1.04	1.70	1.13

Table 3. Rate Rx achieved by the adaptive scheme as compared to Slepian-Wolf bound (Tsukuba)

	D=16		D=22		D=32		D=48	
	R1*	R2**	R1	R2	R1	R2	R1	R2
Fixed Coder	1.40	1.07	1.60	1.10	1.70	1.12	1.90	1.16
Adaptive Coder	1.20	1.03	1.40	1.07	1.50	1.09	1.70	1.13

$^{*}\ R1 = \frac{Rx}{H(CN|PN1,PN2)}$ $^{**}\ R2 = \frac{Rx+H(PN1,PN2)}{H(CN|PN1,PN2)}$

as well as satisfying the Slepain-Wolf bound. Note that in the proposed adaptive coding scheme, we have only considered the CN rate to be close to the information-theoretic bound. However, one of the two parent nodes should be encoded to be able to completely approach the bound in PCP architecture.

4 Conclusion

We proposed an encoder/joint decoder scheme based on adaptive module operation for asymmetric distributed source coding of multi-view images considering the spatial frequency of the area where the encoding algorithm is applied. The CN is encoded without knowledge of PNs. At the encoder the desired quality is controlled by the "D" value, which is used to encode a block in CN image. The output is the module image (i.e., coset image). The decoder performs an inverse module operation by estimating the "D" value using the received encoded CN and the side information provided by geometry compensation using PNs.

The proposed adaptive distributed source coding can approach to the Slepian-Wolf bound by controlling the quality. Furthermore, its performance has a significant improvement in comparison with conventional coding scheme, with smaller average "D" value. We have considered three views and compared the Slepain-Wolf bound for transmission of a view. However, coding of parent node is also needed to satisfy the joint entropy of multi-view image transmission. This issue has been remained as future research. In addition, a suitable compression method for coset image after using the proposed adaptive scheme is required.

References

1. P. Na Bangchang, T. Fujii, M. Tanimoto: Experimental System of Free Viewpoint Television, Vol. 5006, Proc. of IS&T/ SPIE Symposium on Electronic Imaging, Santa Clara, CA, USA (Jan. 2003) 554563
2. D. Slepian and J.K. Wolf: Noiseless Coding of Correlated Information Sources, Vol. IT-19, IEEE Transaction on Information. Theory (July 1973) 471480
3. S.S. Pradhan, and K. Ramchandran: Distributed Source Coding Using Syndromes (DISCUS): Design and Construction, Proc. IEEE Data Compression Conference, Snowbird, UT (March 1999) 157167
4. S.S. Pradhan, and K. Ramchandran: Distributed Source Coding: Symmetric Rates and Application to Sensor Networks, Proc. IEEE Data Compression Conference, Snowbird, UT (March 2000) 363–372
5. X. Wnag and M. Orchard: Design Of Trellis Codes for Source Coding with Side Information at Decoder, Proc. IEEE Data Compression Conference, Snowbird, UT, pp, 361–370, March 2001.
6. J. Kusuma, L. Doherty, K. Ramchandran, Distributed Compression for Sensor Networks, IEEE Signal Processing Society Conference, ICIP (2001)
7. X. Zhu, A. Aaron and B. Girod: Distributed Compression for Large Camera Arrays, Proc. IEEE Workshop on Statistical Signal Processing, SSP-2003, St Louis, Missouri (Sept. 2003)
8. A. Aaron and B. Girod: Compression with Side Information Using Turbo Codes, Proc. IEEE Data Compression Conference, DCC-2002, Snowbird, UT (April 2002)
9. M. P. Tehrani, T. Fujii, M. Tanimoto: A Distributed Source Coding for ITS, Forum on Information Technology, FIT (Sept. 2002) 389–390
10. M. P. Tehrani, T. Fujii, M. Tanimoto: Distributed Source Coding of Multiview Images, Vol. 5308, Proc. of IS&T/ SPIE Symposium on Electronic Imaging, VCIP, San Jose, CA, USA (Jan. 2004) 300–309
11. A. Nakanishi, T. Fujii, T. Kimoto, and M. Tanimoto: Ray-space Data Interpolation by Adaptive Filtering Using Locus of Corresponding Points on Epipolar Plane Image, Vol. 56, The journal of the institute of Image information and Television Engineers, ITE (Aug. 2002) 1321–1327
12. M. Droese, T. Fujii, M. Tanimoto: Ray-Space Interpolation Based on Filtering in Disparity Domain, Proc. 3D Image Conference, Tokyo, Japan, (June 2004) 129–132

Hybrid Multiple Description Video Coding Using SD/MD Switching

Il Koo Kim and Nam Ik Cho

Inst. of New Media and Communications
School of Electrical Eng., Seoul National Univ., Seoul 151-742, Korea
lit2eng@ispl.snu.ac.kr, nicho@snu.ac.kr

Abstract. This paper proposes an algorithm for the robust transmission of video in error prone environment using hybrid multiple description coding (Hybrid MDC) scheme. Multiple description coding (MDC) is more resilient than single description coding (SDC) against severe packet loss rate (PLR) condition. But the excessive redundancy in the MDC degrades the performance in the case of lower PLR. To overcome this problem of MDC, we propose a hybrid MDC method that can switch between SD and MD according to the channel condition. Specifically, the encoder selects SDC for the coding efficiency at low PLR and MDC for the error resilience at high PLR. For the optimal switching of SD/MD the rate-distortion optimization framework is used in this paper. The recursive optimal per-pixel estimate (ROPE) technique is adopted to estimate the accurate decoder distortion at the time of encoding. Experimental results show that proposed Hybrid MDC with SD/MD switching algorithm is more effective than conventional MDC algorithms at low PLR as well as at high PLR.

1 Introduction

Most of information including the multimedia data such as audio, image and video, are split into small size of chunks, so called packets, and transmitted through the network. However, the packet-switching network does not guarantee the end-to-end Quality of Service (QoS). In case of Internet, increment of transmitted data causes traffic jam and packets are discarded due to buffer overflow or long queuing delay at the intermediate nodes of the networks. For increasing the resilience against the channel error under such environment, multiple description coding (MDC) has been studied [1,2,3,4]. In MDC, a source signal is split into several coded streams, which is called descriptions, and each description is transmitted to the decoder through physically or virtually different channels. The main advantage of MDC lies in the fact that the source signal can be reconstructed even when not all the description is received at the receiver. More specifically, the MDC splits the stream into several descriptions which are designed to be correlated with one another. As a result, the missing description can be estimated from the successfully received ones. Hence, the main issues of MDC in signal processing are how to make an effectively structured correlation between the descriptions and how to control the amount of correlation.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 257–264, 2004.

Recently, an MD video coding algorithm based on rate-distortion optimization was developed in [5]. Although three prediction loops are used in this method, residual error signal is not sent to the decoder. Only two side prediction loops are used in the rate-distortion optimization that minimizes the reconstruction error and mismatch. In addition to the error resilience, high coding efficiency is also obtained. Similar algorithm was proposed in [6], which splits a stream into the unbalanced high and low resolution descriptions. To minimize the encoder-decoder mismatch in the rate-distortion optimization, it is required to estimate the exact decoder distortion in rate-distortion sense at the encoding stage. In [7], recursive optimal per-pixel estimate (ROPE) technique [8] is used to estimate the decoder distortion per pixel. ROPE was first introduced to optimally choose the placement of intra-blocks in a one layer encoder. In [7], a single description ROPE (SDC-ROPE) was extended to optimize the placement of intra-blocks within the MD coder, in addition to the optimal allocation of MD redundancy to each block. Multiple description ROPE (MDC-ROPE) gives the performance improvement at 2% packet loss rate or higher. But the performance at below 2% packet loss rate is inferior to the SDC-ROPE case because of unavoidable excessive redundancy, which consists of motion vectors and coding modes.

In this paper, we propose a Hybrid MDC using rate-distortion optimized SD/MD switching scheme, in order to remove the unavoidable excessive redundancy in conventional MDC. In MDC, some redundancy is unavoidable due to the correlations between descriptions. At low packet loss conditions, this redundancy degrades the coding efficiency. To overcome this problem of MDC, we propose a hybrid MDC method that controls the SD/MD switching according to the channel condition. For example, SDC is used for coding efficiency at low PLR and MDC is used for error resilience at high PLR. To control the SD/MD switching in the optimal way, RD optimization framework is used. Lagrange optimization technique selects switching mode that minimizes RD-based cost function, $D+\lambda R$, where R is the actually coded bit rate and D is the estimated distortion.

2 Proposed Hybrid Multiple Description Video Coding

2.1 General Framework

In this paper, we use a simple method to split DCT coefficients generated by the traditional one-layer or single description (SD) encoder. To achieve complete mismatch control, side prediction error signal should be encoded and transmitted. But it was proved in [5] that simple alternation and duplication of DCT coefficients can be a reasonable solution. In this method, rate-distortion optimization is used to achieve good MD performance in this simple condition.

Figure 1 shows the structure of proposed MD video coder when two independent channels are available. The structure and algorithms for multiple channels can be derived from the two-channel scheme. Given the k-th frame F_k and the (k-1)-th reconstructed frame ($\tilde{F}_{0,k-1}$), the encoder estimates the block-based motion vector, MV. The MV is applied to motion predictor, which produces prediction error signal (R_0). The prediction error signal R_0 is split into two

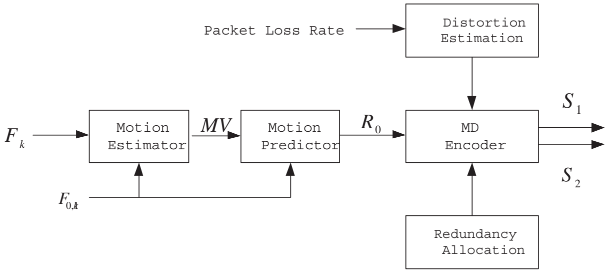


Fig. 1. General framework of proposed Hybrid MDC video coder.

description by the MD encoder and transmitted to the decoder by two independent channels. MD encoder performs optimal split of DCT coefficients and SD/MD switching by estimating the decoder distortion under the given packet loss rate (PLR) and redundancy rate. The optimal split, SD/MD switching and the estimation of decoder distortion algorithm are explained in detail in the following subsections.

2.2 Optimal Split of DCT Coefficients

In order to split a one-layer description into two, we use optimal split of DCT coefficients [9]. Using dynamic programming approach, the optimal split can find a set of coefficients $\tilde{C}_1$ and $\tilde{C}_2$, the elements of which are the duplication or alternation of original coefficients from a set of quantized coefficients $\tilde{C}$. Then, $\tilde{C}_1$ is assigned to the first description S_1 and $\tilde{C}_2$ to the second description S_2. This problem can be formulated as a constrained problem

$$\min_{\tilde{C}_1,\tilde{C}_2} [D(\tilde{C}_1,\tilde{C}_2)] \text{ subject to } \rho(\tilde{C}_1,\tilde{C}_2) \leq \rho_{budget} \tag{1}$$

where D and ρ are distortion and redundancy, respectively and ρ_{budget} is the available redundancy budget. Eq. (1) can be solved by Lagrange optimization method, *i.e.*, it can be solved by minimizing

$$J(\lambda) = D(\tilde{C}_1,\tilde{C}_2) + \lambda\rho(\tilde{C}_1,\tilde{C}_2) \tag{2}$$

where the Lagrange multiplier λ is determined considering the distortion and redundancy budget. If we have the optimal solution up to the j-th coefficients, we can use it to solve the next optimization problem to the k-th coefficient $(k > j)$ by

$$J_k = J_j{}^* + \varDelta J^{op}_{j,k}(k > j) \tag{3}$$

where $\varDelta J^{op}_{j,k}$ is an incremental Lagrange cost using MD operator, *op*. This recursive structure enables the dynamic programming approach. For general k, the minimum cost is found to be

$$J^*_k = \min_{j,op}\{J_j{}^* + \varDelta J^{op}_{j,k}\}, \text{for } j = 0, \cdots, k-1. \tag{4}$$

We can calculate every optimal Lagrange cost J_k^* for $k = 0, \cdots, 63$. In addition to the Lagrange cost, we can find the optimal predecessor j and the optimal operation op_k^* of the k-th coefficient. Then, the k that minimizes J_k^* is denoted by k^*. Clearly, $J_{k^*}^*$ is the minimum Lagrange cost for the whole block. Using the saved information such as predecessor and optimal operation, we backtrack and find the optimized coefficients down to the coefficient 0.

2.3 Estimation of Decoder Distortion

In [9], ideal MD channel environments was assumed, that is, one (or more) channel is totally failed. But this assumption is not valid any more in packet-switching network. Packets can be dropped in the intermediate node without any notice. Hence, we do not know which packet is lost. But we can estimate packet loss rate within an interval of time. Using the packet loss rate (PLR), we estimate the decoder condition, especially the decoder reconstruction distortion at the time of encoding. Considering the PLR, distortion D can be represented as

$$D(\tilde{C}_1, \tilde{C}_2) = \sum_{i=0}^{63} d_n^i \tag{5}$$

where d_n^i is defined as

$$d_n^i = E\{(f_n^i - \tilde{f}_n^i)^2\} = (f_n^i)^2 - 2f_n^i E\{\tilde{f}_n^i\} + E\{(\tilde{f}_n^i)^2\} \tag{6}$$

where f_n^i denote the original value of pixel i in the n-th frame and $\tilde{f}_n^i$ is the reconstructed value at the decoder, possibly after error concealment. For the encoder, $\tilde{f}_n^i$ is a random variable. This concept is first introduced in [8] for the SD video coding as Recursive Optimal per-Pixel Estimate (SDC-ROPE) and extended to MD video coding in [7] (MDC-ROPE). We adopt the MDC-ROPE method with half pixel accuracy to estimate the decoder distortion at the encoder.

2.4 SD/MD Switching

In this subsection, we present the SD/MD switching scheme. Let there be N frames in the video sequence, with M blocks per frame. We wish to select the optimal mode between SD and MD coding by defining the cost for the given mode as follows.

$$mode^* = \min_{mode} E_{i,mode}. \tag{7}$$

Here, $E_{i,mode}$ is a Lagrange cost function defined as

$$E_{i,mode} = D_{i,mode} + \mu R_{i,mode} \tag{8}$$

where μ is Lagrange multiplier, $D_{i,mode}$ is the i-th frame's overall reconstruction distortion and $R_{i,mode}$ is the i-th frame's overall bit rate. μ is proportional to

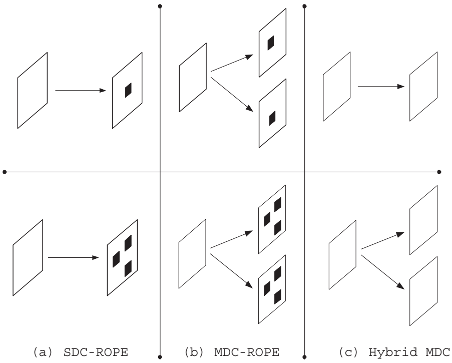


Fig. 2. Illustration of three algorithms used in experiments.

the average of Lagrange multiplier used in the optimal split of DCT coefficients. Specifically, $D_{i,mode}$ and $R_{i,mode}$ is defined as

$$D_{i,mode} = \sum_{j=0}^{M-1} d_{i,j,mode}, \tag{9}$$

$$R_{i,mode} = \sum_{j=0}^{M-1} r_{i,j,mode} \tag{10}$$

where $d_{i,j,mode}$ is the j-th block's reconstruction distortion and $r_{i,j,mode}$ is j-th block's bit rate. $d_{i,j,mode}$ and $r_{i,j,mode}$ are obtained by optimal split of DCT coefficients, which is explained in the previous section.

3 Codec Implementation and Bit Allocation

Proposed MDC codec is implemented based on the ITU-T H.263 video codec. In our coder, different from the conventional SD and MDC video coder, we need to notify SD/MD switching to the decoder. In order to keep the syntax of H.263 bit stream, we make use of PEI (Extra Insertion Information) and PSUPP (Supplemental Enhancement Information) bits in the Picture Header area [10]. The PSUPP data consists of a four-bit function type indication FTYPE, followed by a four-bit parameter data size specification DSIZE. A decoder that receives an unsupported function type indication can discard the function parameter data for that function. Because we use the Extended Function Type to exchange the SD/MD switching information, video decoder that does not support this function type, will discard the SD/MD switching information.

We use TMN8 rate control scheme and Lagrange multiplier selection method to meet the rate constraints. Lagrange multiplier selection is proposed in [11]. By the off-line experiments, the best Lagrange multiplier λ is determined as follows.

$$\lambda = c \cdot (QP)^2. \tag{11}$$

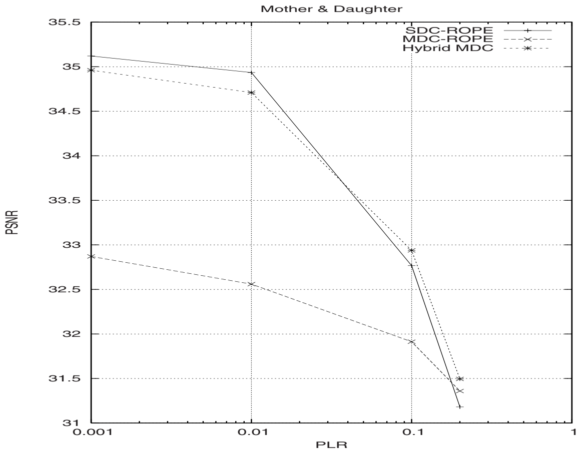


Fig. 3. Packet loss performance for Mother & Daughter sequence at 64kbps

The theoretical justification is also given in the same paper,where the constant c is 0.85 in H.263 video encoder. We use the constant $c = 0.85$ in SD-ROPE. Since the properties of MD video coding is different from SD video coding, we propose a new Lagrange multiplier selection equation as

$$\lambda = c \cdot (QP)^2 \cdot (-log(PLR)). \tag{12}$$

In the MDC, Lagrange multiplier controls error resilience as well as coding efficiency. At high PLR, more redundancy is required to resist the packet loss errors. Thus, PLR should be incorporated into the Lagrange multiplier selection equation. In the modified equation, we use constant $c = 0.15$ in MDC-ROPE and Hybrid MDC, which is obtained by extensive simulation.

4 Experimental Results

We compare the performance of SDC-ROPE, MDC-ROPE and Hybrid MDC using TMN8 rate control, and the modified Lagrange multiplier selection method. Fig. 2 shows the notion of three algorithms used in experiments. In Fig. 2, the upper row is low PLR case and the lower row is high PLR case. Gray colored blocks represent intra blocks. Hybrid MDC uses SD/MD switching according to the channel condition, whereas SDC-ROPE and MDC-ROPE use inter/intra mode switching. We use baseline H.263, and the GOB headers are included. Each frame is split into two descriptions. In case of SD, odd GOBs are assigned into one description and even GOBs are assigned into another description. And intra frame update rate is every 10 frames. We injected 30 different packet loss pattern for each sequence and averaged the results. Fig. 3 and Fig. 4 shows the packet loss performance for Mother & Daughter and Foreman sequences. In case of Mother & Daughter, SDC-ROPE is slightly better than proposed Hybrid

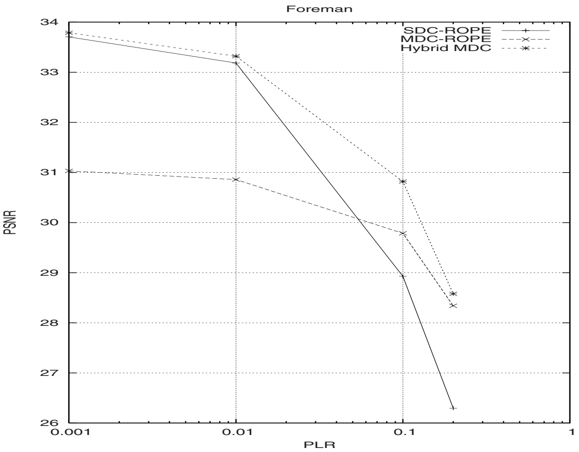


Fig. 4. Packet loss performance for Foreman sequence at 128kbps

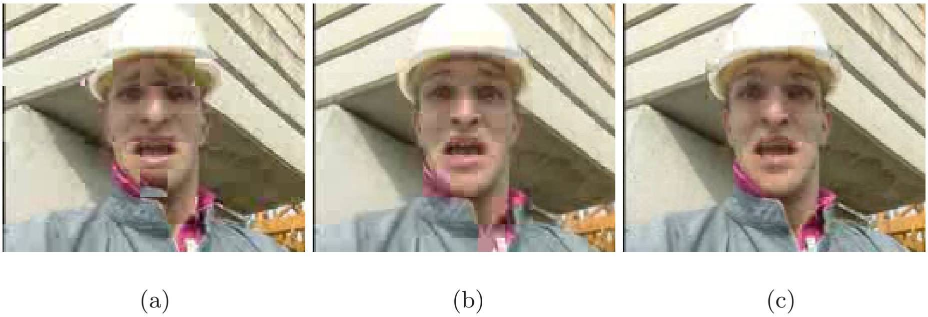


Fig. 5. Comparison of decoded picture quality, Foreman, 128kbps. (a) SDC-ROPE (b) MDC-ROPE (c) Hybrid MDC

MDC in low PLR because of low motion property of the sequence and SD/MD switching information. But in case of Foreman, proposed Hybrid MDC is more effective than the conventional SDC-ROPE and MDC-ROPE algorithm at low PLR as well as at high PLR. Fig. 5 shows the decoded pictures of Foreman for the subjective comparison. It can be observed that the SDC-ROPE suffers from annoying blocking artifact due to error concealment, whereas the MDC-ROPE and Hybrid MDC have no blocking artifact. But, decoded picture quality of MDC-ROPE is lower than the Hybrid MDC because of excessive redundancy.

5 Conclusions

An algorithm for the robust transmission of video in error prone environment is proposed, which uses multiple description coding (MDC) with SD/MD switching. Specifically, SD/MD switching scheme is employed in order to overcome the excessive redundancy problem. SDC is used for the coding efficiency at low PLR condition and MDC is used for the error resilience at high PLR. To control the SD/MD switching in the optimal way, the RD optimization framework is used. Lagrange optimization technique minimizes the RD-based cost function, $D+\lambda R$, where R is the actually coded bit rate and D is the estimated distortion. The recursive optimal per-pixel estimate (ROPE) technique is adopted to estimate the accurate decoder distortion at the encoding time. Simulation results show that proposed SD/MD switching algorithm is more effective than the conventional MDC algorithms at low PLR conditions as well as at high PLR.

References

1. V. Goyal, "Multiple description coding : Compression meets the network," *IEEE Signal Processing Mag.*, vol. 18, pp. 74–93, Sept. 2001
2. A. Ingle and V. Vaishampayan, "DPCM system design for diversity systems with applications to packetized speech," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 48–57, Jan. 1995
3. Y. Wang, M. Orchard, and A. Reibman, "Multiple description image coding for noisy channels by pairing transform coefficients," in *Proc. First Workshop on Multimedia Signal Processing*, June 1997, pp. 419–424
4. A. Reibman, H. Jafarkhani, Y. Wang, M. Orchard, and R. Puri, "Multiple description coding for video using motion compensated prediction," in *Proc. Int. Conf. Image Processing*, Oct. 1999, pp. 837–841
5. A. Reibman, H. Jafarkhani, Y. Wang, and M. Orchard, "Multiple description video using rate-distortion splitting," in *Proc. Int. Conf. Image Processing*, Oct. 2001, pp. 978–981
6. D. Comas, R. Singh, and A. Ortega, "Rate-distortion optimization in a robust video transmission based on unbalanced multiple description coding," in *Proc. of 2001 IEEE Fourth Workshop on Multimedia Signal Processing*, Oct. 2001, pp. 581–586
7. A. Reibman, "Optimizing multiple description video coders in a packet loss environment," in *Proc. Int. Packet Video Workshop*, Mar. 2002
8. R. Zhang, S. Regunathan, and K. Rose, "Video coding with optimal Inter/Intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 966–976, June 2000
9. I. K. Kim and N. I. Cho, "Error resilient video coding using optimal multiple description of dct coefficients," in *Proc. Int. Conf. Image Processing*, Sept. 2003
10. Z. Jia, K. Tang, and H. Cui, "A H.263 compatible error resilient video coder," in *Proc. Int. Conf. WCC-ICCT 2000*, Aug. 2000, pp. 1157–1160
11. T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proc. Int. Conf. Image Processing*, Oct. 2001, pp. 542–545

Clusters-Based Distributed Streaming Services with Fault-Tolerant Schemes

Xiaofei Liao and Hai Jin

Cluster and Grid Computing Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
{xfliao,hjin}@hust.edu.cn

Abstract. Distributed streaming servers based on clusters with multi-components have several failure points, which will destroy the systems or degrades the quality of services. How to provide reliable streaming services is a very important issue. Fault-tolerant schemes must be provided for video servers to improve system reliability and availability. This paper presents novel distributed media streaming services based on cluster architectures with fault-tolerant schemes, including video-on-demand services and video recording services. The new system eliminates three kinds of failure points and addresses the server and network failures, not only the disk or file failures. Experiment results have been proven that the system has better performance.

1 Introduction

Designing a good VOD (Video-on-demand) system is a big challenge. It is very necessary to find a balance between two parameters, scalability and reliability. When we design massive streaming services based on clusters, the system consists of very complex components and layered structures. Although the system has good scalability, it is not very reliable and has multiple failure points.

Many researchers propose different designs of video servers [2] based on clusters. But these systems are not involved with fault tolerant schemes. How to provide reliable streaming services is a very important issue. The issue involves two aspects. One is about the disk or file failure, called data invalidations; the other one is about the server or network failure, called connections invalidations. Most people study the fault tolerant schemes [4] with the focus on the disk or file failures, such as redundant data placement policies, reliable disk arrays. In contrast to redundant disks or files schemes, how to keep the services stable and continuous transparently in case of server or network failures is a big challenge.

In this paper we propose a novel distributed streaming server based on clusters to provide video-on-demand services and recording services. We propose different fault-tolerant schemes to eliminate the corresponding failure points and improve the system reliability without degradation of scalability.

The rest of this paper is organized as follows: section 2 overviews the related works regarding the fault tolerant schemes of cluster video servers; section 3 presents the system architecture; section 4 discusses the fault tolerant schemes;

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 265–272, 2004.

section 5 gives the experiment results and some discussions; finally we conclude in section 6.

2 Related Works

Many researchers put their emphases on providing streaming services with high reliability. Most people would like to study the disk or file failures [6][8]. Lee [6] presents and analyzes a fault tolerant mechanism based on inter-node striping and erasure correction codes to tackle this challenge. By formulating the system reliability by using a Markov chain model, Lee obtains insights into the feasible operating region of the system, such as the amount of redundancy required and the node-level reliability that can be tolerated.

In [8], a general scheme for accomplishing both fault tolerance and high availability in a video server is proposed. The tradeoffs between memory buffer space and redundant disk storage space are discussed. A video allocation problem based on the proposed scheme is formulated and heuristic algorithms to solve this problem are presented. It is obvious that the paper focuses on the disk storage and only involves the server failures.

Other researchers take the server or network failures into consideration [1][3]. In [1], authors provide a new scheme, which can support smooth migration of clients from one server to another. The system uses a group communication scheme to loosely coordinate the participating servers to agree upon client migration and to allow one server to take over another server client in case of server or network failures. But the video transmission may stop for a short period, frames may arrive twice, or may arrive out of order.

In [3], three fault tolerant video streaming models are discussed. These models guarantee continuous streaming to clients despite of server failures, while utilizing very low network bandwidth and a small client buffer. But the scheme is achieved by exploiting the special characteristics of the APEG codec and Motion-JPEG codec, which needs many system resources.

3 Cluster-Based Distributed Streaming Servers

The architecture of our distributed streaming server based on clusters is showed in Fig. 1. There are three layers in this system. Layer 1 is a virtual server, which is the single entry point. Atlayer 2, control servers are responsible for admission control and parsing clients streaming requests into sub-tasks. Data servers and record servers are at layer 3.

Under the control of record servers, record terminals will catch analog video and audio signals, which will be sent to video capture cards on record servers. Record servers then compress these signals into digital data according to media formats standards with the help of video capture cards. All these media data are formatted as RTP (Real-time Transmission Protocol) packets according to Owl scheme [5] and are sent to data servers with the instruction of control

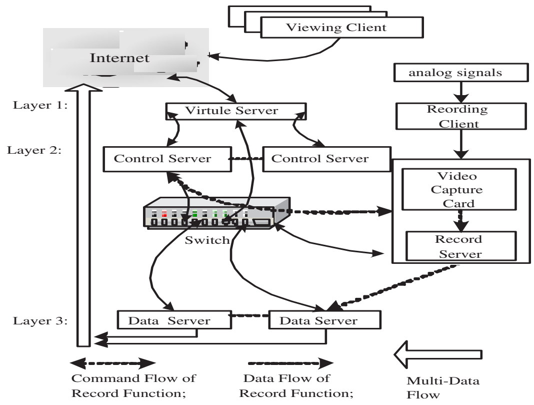


Fig. 1. Architecture of Cluster-based Media Streaming Server with Video-on-Demand Services and Video Recording Services.

servers according to D-Record protocol [7]. Control servers are the most important nodes. When providing video-on-demand services, data servers are running RTP servers.

There are two data flows in the system: command flow and media data flow. Command flow includes media requesting commands and recording commands. There are two types of media data. One is called recorded media data and another one is called on-demand media data. Media data being recorded are flowing inside the clustered systems, transferred from recording servers to distributed data servers. Contrarily, on-demand media data are sent to viewing clients directly via attached networks.

From the architecture, we find out the system has two obvious failure points. One is from recording sessions and the other is caused from on-demand sessions. How to keep these two services stable and reliable is the main task of this paper.

4 Fault-Tolerant Schemes

There are three failure points in the system: the control server failure, the data server failure point and the recording server failure.

4.1 Fault Tolerant: When Data Servers Crash

There are two problems caused from the invalidation of data servers: the normal on-demand sessions will be destroyed and the record sessions will be demolished.

Consider the recovery of normal video-on-demand sessions at first. Every data server involved in one video-on-demand session must report its own status to the control server via PING-PONG mechanism. When the control server cannot receive one message from one data server, it will do the following works: select new data server from backup lists; the information at the break point are recorded;

the control server rebuilds a new request according to the information and send it to the new data server. The new request is according to RTP/UDP protocol.

Then we consider the recovery of the record sessions. Suppose the record server never crash, the protocol D-Record can provide fault tolerant schemes to ensure that the recording functions can work well when one or several data servers crash. When record servers prepare to send media data out, there will be a data server group to receive current clip's media data. An effective method to transmit media data is under the control of an application-layer multicast group. Any failed data server, will be quit out from this group.

It is very important to set a threshold of the timeout value *TotalTimeOut* and the number of timeout *TotalTimes* when control servers begin to wait for the data servers response to the command Record-Startup. These two parameters limit the times of failures and save systems resources. When the timer is up to the threshold value and there are no full responses from current data server group, control servers will delete the data servers who have no responses and select another one. When *TotalTimes* equals zero, current record session failes and cannot be recovered.

4.2 Fault Tolerant: When Control Servers Crash

Each control server maintains two long-connections and two types of sub-tasks lists. These two long-connections, viewing clients connections based on RTSP protocol and record servers connections based on D-Record protocol, are all based on TCP. The viewing sub-tasks lists are be used to dispatch the jobs for the data servers should do and are created temporarily after the requests from viewing clients being parsed. The recording sub-tasks lists are created under the control of D-Record and are used to instruct that to which data servers the record servers should transmit the recording media data. The fault-tolerant schemes of control servers should keep these connections and reconstruct the same sub-tasks lists.

It is well known that TCP based connections are difficult to be recovered for their complex status information and accurate sequences numbers. But if we do not consider the universality of the schemes, it is very easy to buildup a fault-tolerant scheme based on clients and the record servers.

First, we consider the connections from viewing clients. Figure 2 shows the concept. Main work to persist these connections will be finished on clients and can be divided into the following parts: a network status monitor, a current playing point recorder, and a connections constructor. The network status monitor will watch the network and make a feedback when the net does not work. When the network monitor finds out that the network does not work, it can point out errors.

The second part, a current playing point recorder, will write down the playing information about the current requests, which include the following parts: requested movie name, current time point and the service address. When network status monitor finds out net is abnormal, systems will stop the current sessions,

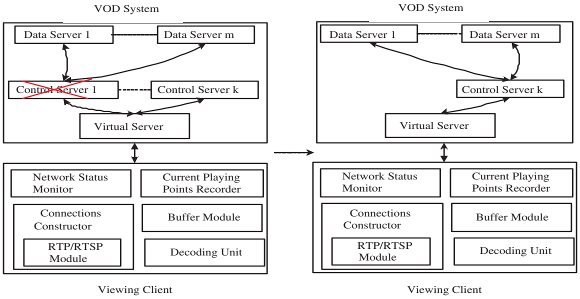


Fig. 2. Recovery of Viewing Connections Based on Viewing Clients.

clear all buffers and reconstruct the connections according to the current playing points information. The VOD server will accept a new request and dispatch the new request to a new control server. It is very necessary to set a threshold to limit the times of reconstruction operations. Video-on-demand services are recovered from the last playing point.

Other very important connections are constructed between record servers and control servers. Figure 3 describes the main idea.

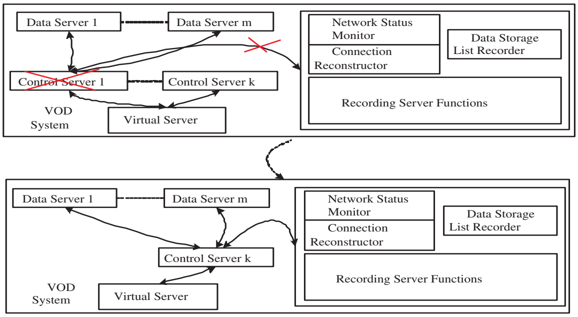


Fig. 3. Recovery of Recording Connections Based on Recording Servers.

At first, when record servers build up connections with control servers, they need to request information from control servers, including the subtask lists of media data storage recorded. With the help of subtask lists, record servers can restart new record sessions to recover the old sessions.

Secondly, control servers must report their own status information to the network status monitor of the record server and tell the current data server ID to data storage list recorder of the record server periodically. With this method, record servers can know the running status of control servers.

At the third step, if record servers find out that one control server does not work, they will start up a new connection with a new control server and send some useful information to the new control server. When the new control server receives the requests from record servers, it builds its own subtask lists.

5 Performance Evaluation

There are two types of hand-over delay in the fault-tolerant system. One is from that the data transmission operation is diverted from one failed data server to normal one; the other one is from that the crashed control servers must be replaced by normal control servers. The expriment environment is as following: two control server, two record servers, eight data servers, and one virtual server. There is one simulated video capture card with 8 input channels in each record server.

5.1 Hand-over Delay Among Data Servers and Control Servers

This experiment tries to find out how long hand-over delay to recover the two kinds of services is when data servers crash and control servers crash. Figure 4 shows the relationships between hand-over delay to recover the video-on-demand services and concurrent streaming number when one data server crashes. And Fig. 5 is about the collapse of one control server.

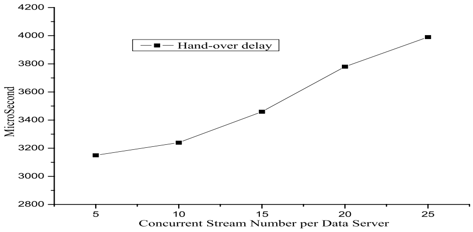


Fig. 4. Relationships between Hand-Over Delay and Concurrent Stream Number When One Data Server Crashes.

In Fig. 4 and Fig. 5, there are five situations with different concurrent stream number each data server, such as 5, 10, 15, 20, and 25. In each situation, we fix the concurrent stream number and shutdown one data server randomly. The control server will find out the failure and then communicates with the backup data server to wait for responses from the new data server. From figures, the system load is higher with the increment of the concurrent stream number, and the hand-over delay is bigger, but in a reasonable range.

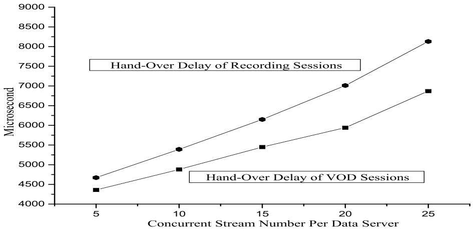


Fig. 5. Relationships between Hand-Over Delay and Concurrent Stream Number When One Control Server Crashes.

5.2 Services Reliability and System Performances

The objectives of clip replication are twofold: fault tolerance and load balancing. The following questions arise: how to get a tradeoff between the service reliability and the storage cost. To explore these issues, we study three different replicate patterns: single-replica, each movie has only one replica; double-replica, each movie has two replicas; variable-replica 1, each one of the top 10 (most frequently accessed) movies has replica degree of 3, and the remaining movies have replica degree of 2.

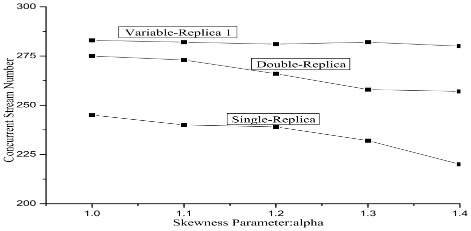


Fig. 6. Relationships between Skewness Parameter and Concurrent Stream Number with Three Situations: Single-Replica, Double-Replica and Variable-Replica 1.

Rather than scheduling between 2 replicas, it chooses the replica with minimum load among n (bigger than 1) replicas to assign new tasks. The experiment results of performance vs. accessing skewness degree are presented in Fig.6 (In all cases we set the number of storage nodes to 8, number of clips to 10, and number of movies to 200).From single-replica scheme to double-replica scheme, there is an obvious improvement in system performance. This is attributed to the load balancing effect by the scheduling algorithm. From double-replica scheme to variable-replica 1 scheme, there is only slight performance improvement, which suggests that the addition of replicas on the base of double-replica scheme only brings quite limited benefit.

In single-replica scheme, the system performance will decline with the increasing of skewness parameter, but shows little degradation when replicate degrees for the movies, especially those popular ones, are raised above 1. This means the proposed scheduling algorithm has the expected effect of smoothing the variability of movie accesses and is fairly robust when the access pattern is changed.

6 Conclusions

In this paper, we propose a new fault-tolerant mechanism for cluster-based streaming server with video-on-demand services and video recording services. According to the experiment results, the management cost of the fault-tolerant scheme is not high and the performance is satisfied.

Acknowledgements. This paper is supported by National 863 Hi-Tech Research and Development Project under grant No.2002AA1Z2102. The final revision has got help from Xin Zhu, at Aizu university.

References

1. T. Anker, D. Dolev, and I. Keidar, "Fault tolerant video on demand services", *Proceedings of 19th IEEE International Conference on Distributed Computing Systems*, 1999, pp. 244–252
2. C. Shahabi, R. Zimmermann, K. Fu, and S.-Y. D. Yao, "Yima: a second-generation continuous media server", *Computer,* Vol.35, No.6, pp. 56–62, June 2002
3. R. Friedman, L. Baram, and S. Abarbanel, "Fault-tolerant multi-server video-on-demand service", *Proceedings of International Parallel and Distributed Processing Symposium*, 2003, pp. 70–77
4. H. M. Vin, P. J. Shenoy, and S. Rao, "Efficient Failure Recovery in Multi-Disk Multimedia Servers", *Proceeding of 25th International Symposium on Fault Tolerant Computing Digest*, 1995
5. H. Jin and X. Liao, "Owl: A New Multimedia Data Splitting Scheme Based on Cluster Video Server", *Proceedings of EUROMICRO Conference*, pp. 144–151, 2002
6. J. Y. B. Lee and R. W. T. Leung, "Design and analysis of a fault-tolerant mechanism for a server-less video-on-demand system", *Proceedings of Ninth International Conference on Parallel and Distributed Systems*, 2002, pp. 489–494
7. X. Liao and H. Jin, "A new cluster-based distributed video record server", *Proceedings of International Conference on Multimedia and Expo,* Vol.3, pp. 249–252, 2003
8. Y. Wang and D. H. C. Du, "On Providing Highly Available Fault-Tolerant Video-on-demand Services", *Proceeding of IEEE International Conference on Multimedia Computing and Systems*, 1998, pp. 76–89

Towards SMIL Document Analysis Using an Algebraic Time Net

A. Abdelli and M. Daoudi

LSI Laboratory
Computer Science Institute of U.S.T.H.B. University
BP 32, El-Alia, Algiers, Algeria
karim.abdelli@wissal.dz

Abstract. This paper presents a formal approach for the design of SMIL documents into a language based on time Petri net (ATN: Algebraic Time Net). The obtained ATN's model is then translated to an equivalent time graph, which describes all possible scenario, and gives for each multimedia sub sequence its minimum and maximum durations. This can be used to guarantee a consistent presentation for the client player and to improve the multimedia server's QoS, by implementing a dynamic scheduling politics.

1 Introduction

In recent years, several papers were focused in the development of the interactive multimedia documents based on the W3C norm: Synchronized Multimedia Integration Language (SMIL) [1]. These works were mainly based on the edition requirements specifications and on the synchronization constraints [2,3]. Others proposed solutions in semantic checking problems and in improving quality of service in distributing such documents [5,6,7]. We propose in this paper an approach, for parsing SMIL document into a model based on Time Petri Net (ATN: Algebraic Time Net) [8,9]. Once the equivalent ATN frame (modelling the SMIL document) is obtained, it is then derived into a reachability graph called "time graph" [4,10]. The generated graph is more compact then the one obtained in [7, 11]. Furthermore, during its construction, the algorithm [10] computes the minimum and maximum durations for each multimedia sub sequence. This allows us then to realize both quantitative and qualitative analysis. The time graph can be used as a deciding tool when editing or executing a presentation and as a predicting tool in improving the quality of service at multimedia server level. This paper is organised as follows: The section two introduces slightly the SMIL language. In section three, we give the adopted approach to translate a SMIL document into an Algebraic Time Net frame. The following section motivates the need to perform this translation using the deduced time graph from the obtained ATN frame to analyse the multimedia document consistency. In the last section, we show how to use the information contained in the time graph to implement a technique for QoS predictive management when distributing the SMIL presentation.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 273–281, 2004.

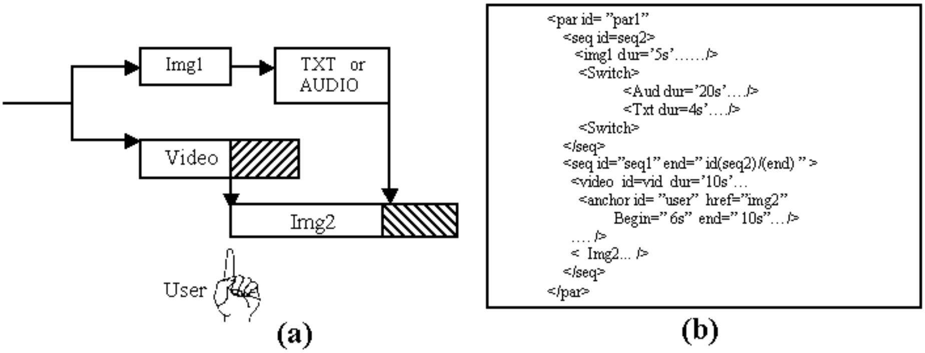
2 Modelling a Multimedia Document with SMIL

SMIL(Synchronized Multimedia Integration Language) [1] is a declarative language based on XML which allows to describe the spatial and temporal aspects of different multimedia objects presentation. It uses the following constructor elements:

- A basic multimedia object "O": It can be either a video, an audio, an image, or a text. It is mainly characterised by its beginning D(O), its end F(O) and its duration Dr(O). D(O) and F(O) must be temporal values or events synchronisation with other multimedia elements. (e.g. D(O):=F(C) means the O object begins when the C object is ending). Note that for each basic multimedia object, we associate two events, its begin event noted B(O) and its end event noted E(O)
- An Anchor or an a element: It specifies an URL link into a basic multimedia object. Only the "Anchor" element can be characterised by temporal values specifying its beginning and its end
- $< par >$ and $< seq >$ elements: They support the same attributes as a basic multimedia object. Note that the element $< par >$ can specify the ending mode synchronisation using the *endsync* attribute. Three values for *endsync* are possible: $first$: expresses that the end of one of its children implies the end of $< par >$, *Last* : $(Defaultvalue)$ the end of the last child implies the end of $< par >$, $id(E)$: the end of $< par >$ is decided by the end of the child, identified by $id(E)$.

Note: For readability raisons, we only consider here the main SMIL.1.0 operators, given above.

Example:



```
<par id= "par1"
  <seq id=seq2>
    <img1 dur='5s'....../>
     <Switch>
            <Aud dur='20s'..../>
            <Txt dur=4s'..../>
     <Switch>
  </seq>
  <seq id="seq1" end=" id(seq2)/(end) " >
   <video  id=vid  dur='10s'...
     <anchor id= "user"  href="img2"
          Begin=" 6s"  end=" 10s"... />
   .... />
   <  Img2... />
  </seq>
</par>
```

Fig. 1. A Multimedia representation and its SMIL document

Consider the scenario of a SMIL document illustrated in Figure 1(a,b) which consists in a sequence seq_1 of a video clip (Vid) followed by an image (Img_2). This sequence must be presented simultaneously with an other image (Img_1)

followed by some related information. This information corresponds to an element of the SMIL operator "switch" such as an audio segment (Aud) or a text (Txt). The end of the presentation of the second sequence seq2 is determined by the end of (Txt) or ($Audio$) (an event synchronisation having the form $F(Seq_1) := F(Seq_2)$). Assume that the durations of Img_1, Aud, Txt and Vid are, respectively, 5 20, 4 and 10 seconds, and that the Img_2 media object does not have an explicit duration. Note that the media Vid can be ended by an external event representing an user interaction. This should occur between 6 and 10 seconds after the beginning of the media Vid.

3 Specifying a SMIL Document Using an ATN Frame

In our approach, we suppose that the SMIL document is syntaxically correct and is edited by an authoring system [2]. It follows then, that it can not exist any ambiguities between the multimedia element temporal attributes. (e.g. If the begin and the end values of the O object are defined, then its duration must take an undefined value or value equal to $F(O) - D(O)$). Our approach translates first, the SMIL document into an *abstractTree*, which describes its logical organisation, in term of composite elements and basic elements. The intermediate tree nodes represent the $< par >$, $< seq >$, $< switch >$ or the $< a >$ constructors. The final tree nodes represent the basic multimedia objects or the *Anchor* element. For each tree node, we save the object identifier $id(O)$, its type $typ(O)$, its duration $Dr(O)$, its beginning $D(O)$, its end $F(O)$, and the set of its ending events $Evf(O)$. Moreover, fort each *par* constructor node, we save its *Endsync* value and the set of the its children event synchronisations $Evs(par)$. Finally, the root of the abstract tree contains the SMIL document name and the set of its ending events.

Remark: A specified event synchronisation in $D(O)$ or $F(O)$ can't deal with elements belonging to the same *Switch* or *Seq* objects. Only the *par* constructor offers the possibility to specify event synchronisations between its children.

The Figure 2(a) presents the abstract tree corresponding to the SMIL document given in figure 1. See that in the tree root, the set $Evf = \{E(Txt), E(Img_2), E(Aud)\}$ indicates the ending events set of the SMIL presentation. In the par_1 node, the set Evs contains the event synchronisation $\{E(seq_1) := E(seq_2)\}$, between the ends of seq_1 and seq_2.

The obtained abstract tree is then parsed into an ATN's frame [9]. The building process is done progressively, using the constructing rules given in Figure 3, going through the abstract tree downwards first and then left to right. The ATN frame allows to model a real time presentation using an algebraic expression which combines algebraic operators with units. Each unit describes here a basic multimedia object which is modelled by a Time Petri Net [8]. To capture the SMIL constructor semantics, we use the next four operators:

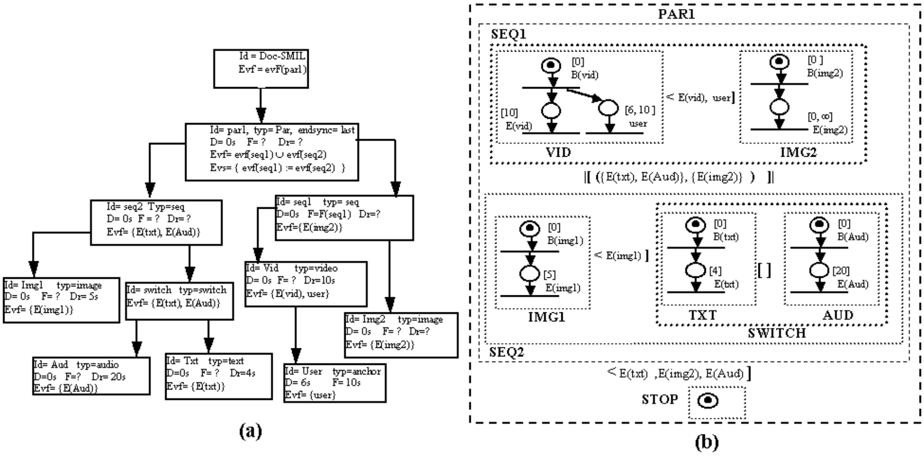


Fig. 2. The SMIL document abstract tree and its Algebraic Time Net

- The choice operator $U_1[\,]U_2$ expresses the non deterministic choice between U_1 and U_2 units. The choice is taken when one of the two units executes an initial events
- The pre-emption operator $U_1 < ev_1..ev_n]U_2$ specifies that the unit U_1 is pre-empted by the unit U_2 if U_1 executes one of $\{ev_1..ev_n\}$ events
- The interruption operator $U_1[ev_1..ev_n > U_2$: specifies that the unit U_1 can be interrupted by the unit U_2, as soon as U_2 realises one of its initial events. In other hand, the interruption mechanism is inhibited if U_1 executes one of $\{ev_1..ev_n\}$ events. In this case the unit U_1 can not be longer interrupted by U_2
- The parallel composition with rendezvous $U_1|[R_1..R_n]|U_2$ denotes the parallel composition of U_1 and U_2 units executing $R_1..R_n$ rendezvous. A rendezvous $R?(\{evm_1..evm_n\}, \{ev_1..ev_m\})$ expresses that the realisation of an event in $\{ev_1..ev_m\}$ is done if a master event in $\{evm_1..evm_n\}$ can be hold at the same time

The Figure 2(b) represents the corresponding ATN's frame of the given SMIL document in Figure 1. Each multimedia object is modelled by an unit. The Seq_2 configuration begins with the Img_1 unit which can be pre-empted by the Switch construction on $\{E(Img_1)\}$ event. The Switch box specifies the choice between the Txt and Aud units. The Seq_2 configuration, begins with the Vid unit which can be pre-empted by the Img_2 unit on $\{E(Vid), User\}$ events. The Seq_1 and Seq_2 sequences are composing in parallel, and are synchronising in $(\{E(Txt), E(Aud)\}, \{E(Img_2\})$ rendezvous. Finally, the Par_1 composition is pre-empted by the Stop unit if one of the $\{E(txt), E(Img_2), E(Aud)\}$ ending events is executed. The ATN building process is done hierarchically, following the rules given in Figure 3. The rule in Figure 3(a), represents a Time Petri Net modelling a basic multimedia object, which is specified by its beginning and

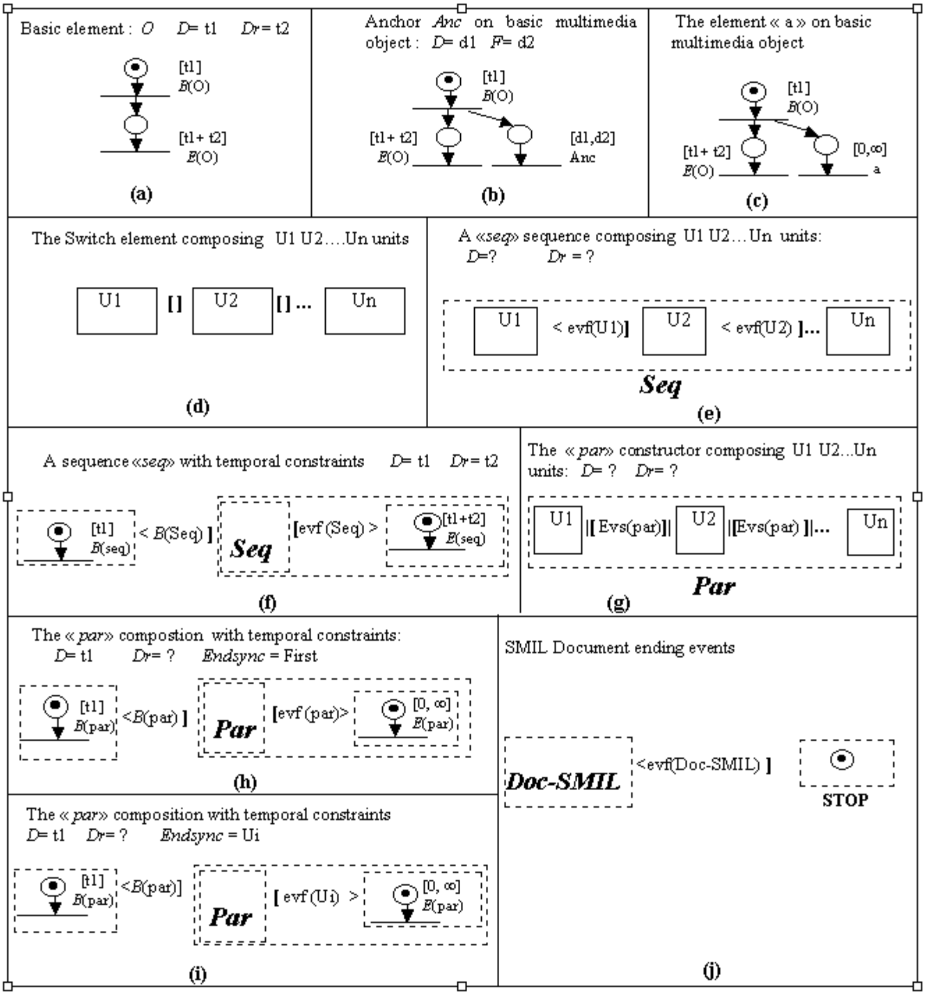

Basic element : O D= t1 Dr= t2
[t1]
B(O)
[t1+ t2]
E(O)
(a)
Anchor Anc on basic multimedia object : D= d1 F= d2
[t1]
B(O)
[t1+ t2]
E(O)
[d1,d2]
Anc
(b)
The element « a » on basic multimedia object
[t1]
B(O)
[t1+ t2]
E(O)
[0,∞]
a
(c)
The Switch element composing U1 U2….Un units
U1 [] U2 []... Un
(d)
A «seq» sequence composing U1 U2…Un units:
D=? Dr = ?
U1 < evf(U1)] U2 < evf(U2)]... Un
Seq
(e)
A sequence «seq» with temporal constraints D= t1 Dr= t2
[t1]
B(seq)
< B(Seq)]
Seq
[evf (Seq) >
[t1+t2]
E(seq)
(f)
The « par» constructor composing U1 U2...Un units: D= ? Dr= ?
U1 |[Evs(par)]| U2 |[Evs(par)]|... Un
Par
(g)
The « par» compostion with temporal constraints:
D= t1 Dr= ? Endsync = First
[t1]
B(par)
<B(par)]
Par
[evf (par)>
[0, ∞]
E(par)
(h)
The « par» composition with temporal constraints
D= t1 Dr= ? Endsync = Ui
[t1]
B(par)
<B(par)]
Par
[evf (Ui) >
[0, ∞]
E(par)
(i)
SMIL Document ending events
Doc-SMIL
<evf(Doc-SMIL)]
STOP
(j)

Fig. 3. ATN frame building rules

ending events. The rules in Figure 3(b,c) model the basic multimedia object interaction with an Anchor or the *a* element. The configuration in Figure 3(d) represents the *Switch* element using the choice operator. Through the rule in Figure 3(e), we model a sequence $(U_1..U_n)$ with no specified temporal constraints on its beginning and its end. If it is the case (see Figure 3(f)), the sequence *seq* is surrounded by two events specifying its begin and its end, with their associated time constraints. The sequence *seq* can be forced to stop if its ending time is elapsed (execution of the $E(seq)$ event) or it finishes normally in respect to its time duration (execution of one of $evf(seq)$) final events). The configuration shown in Figure 3(g) models the *par* constructor using the parallel composition of $U_1..U_n$ units and synchronizing on $evs(par)$ rendezvous. For each event synchronisation from $evs(par)$ having the form $\{ev_i := ev_i'\}$ corresponds the rendezvous $R_i = (\{ev_i'\}, \{evi\})$ in the parallel composition operator $|[R_1...R_m]|$. The configurations in Figure 3(h,i) model the cases where the Endsync attribute don't take the default value *last*. Note that in SMIL1.0 specification, the *Endsync* value is not interpreted if the duration or the end of the *Par* element is specified. Finally, the rule given in Figure 3(j) allows to achieve the ATN building process, specifying the events set $evf(document)$ on which the SMIL multimedia document ends. The execution of one of these events ends the document presentation.

4 Consistency Analysis and Scheduling Using the Time Graph

The ATN reachability graph called time graph is generated using an approach developed in [10]. Each graph node contains temporal constraints defined on the variables k_i which determines the consistency value domains. The variable k_i represents, here the elapsed time between the $(i-1)^{th}$ and the $(i)^{th}$ nodes. When generating the graph, the algorithm computes also for each terminal node (n) the tables $D_n(i,j)$ which gives $(if(i < j))$ the maximum and $(if(i > j))$ the opposite value of the minimum elapsed times between the nodes (i) and (j). Each branch graph is labelled with the set of parallel events which should occur at the same time, if the temporal constraints defined on the branch ending node are satisfied. The time graph obtained from a SMIL document ATN's frame gives all possible multimedia scenarios. Also, it is more compact then the one obtained in [11]. The generating algorithm [10] uses a bisimulation which contracts the graph and reduce the state number explosion. This is done by regrouping in one branch all the parallel events that could occur at the same times. Furthermore, the computed tables D_n allow us to think on either quantitative or qualitative analysis of the SMIL document. The formal verification of SMIL document consistency is treated in [11]. A document is said to be consistent if the action characterising the beginning of the presentation is necessarily followed (and this for every path considered in the graph) by an action characterising its end. In other words, the non consistency of a multimedia document could be defined first, by identifying the origins of its inconsistencies in the temporal scenario, and then check where they can be corrected by temporal formatting. In Figure 4 we give the time graph describing the SMIL document already presented in Figure 1. We can see that the SMIL document is consistent since all the graph paths lead to the occurrence of one of $\{E(img_2), E(Txt), E(Aud)\}$ events.

In conclusion, the time graph could be used at two levels:

- At editing level: it allows to check the SMIL document temporal consistency. If the document is not consistent, one has to format it until obtaining a consistent graph called *scheduling graph*
- At the Player level: If we need to have a consistent presentation from a non consistent document, a consistent graph is extracted from the inconsistent one when removing the inconsistent branches. Then the player will use the

D_{42}	0	1	22	33	42
0	0	0	5	10	25
1	0	0	5	10	25
22	-5	-5	0	5	20
33	-6	-6	-1	0	19
42	-25	-25	-20	-15	0

D_{43}	0	1	22	34	43
0	0	0	5	10	25
1	0	0	5	10	25
22	-5	-5	0	5	20
34	-10	-10	-5	0	15
43	-25	-25	-20	-15	0

D_{41}	0	1	21	31	41
0	0	0	5	9	9
1	0	0	0	9	9
21	-5	-5	0	4	4
31	-6	-6	-1	0	3
41	-9	-9	-4	0	0

D_{32}	0	1	21	32
0	0	0	5	9
1	0	0	5	9
21	-5	-5	0	4
32	-9	-9	-4	0

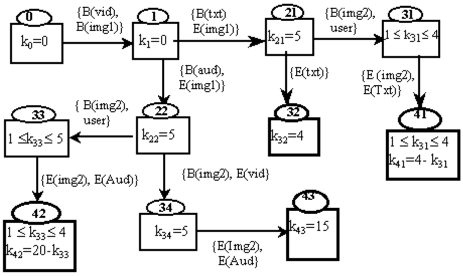


Fig. 4. The SMIL document time graph an its D_n tables

obtained graph as a deciding tool, to only present the consistent multimedia scenarios. The D_n tables can be used here to schedule an appropriate multimedia sub sequence presentation, considering the minimum and the maximum durations of each one.

5 QoS Predictive Management in the SMIL Document Presentation

The multimedia server has to make sure that the multimedia stream temporal order is satisfied when answering the client request. Within networks a server may need to service many clients at the same time. So, during a presentation a client may have to wait indefinitely before receiving the next multimedia data stream. This happens because client requests are queued by a server before they can be serviced. This results on a loss of synchronization between the different media objects which are required by a client for its presentation. In this case, we propose a technique by which the server hosting the SMIL document pages, informs in advance the different multimedia servers of the required media objects in the future for their client player presentation. This information can be used by each multimedia server to schedule and optimise the delivery of multimedia data streams. This technique is described in Figure 5 and follows the next steps:

1. The client sends a request to the host server, hosting the SMIL XML pages
2. When receiving the request, the server loads the XML code, and retrieves its translated time graph from a storage Data base. For each required component E_i in the presentation, it extracts from the tables Dn (when considering all the possible paths in the graph) the minimum elapsed time between the beginning of both document and E_i component, Time-begin(E_i) and the maximum elapsed time between the beginning and the end of the component, Time-max(E_i). (E.g. For the Vid media object, we obtain: Time-begin(Vid) = $0s$ and Time-max(E_i) = $10s$)

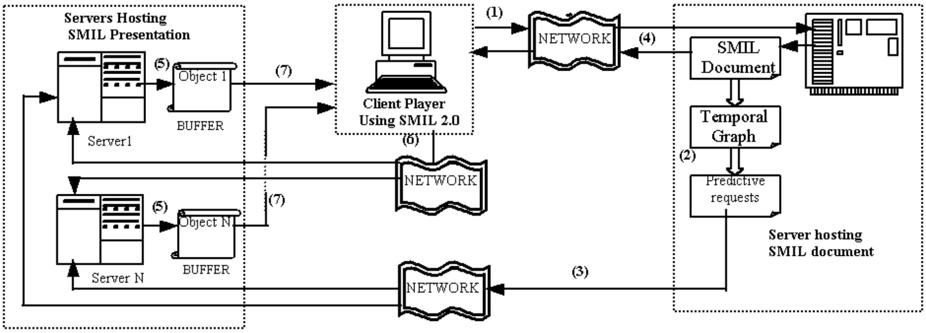


Fig. 5. QoS Predictive Management

3. The Web server sends a special predictive request to each multimedia server, (hosting the component E_i which uses streaming), and asks him to load Time-max(E_i) of the E_i component which will be requested in at least Time-begin(E_i) seconds by the identified client
4. The web server sends the document SMIL XML code to the player client
5. When receiving the predictive request, the component E_i multimedia server host deals first with the urgent request (value of Time-begin(E_i)), loading in a buffer the asked part of the E_i component which length of time is Time-max(E_i)
6. When receiving the XML code of the SMIL document, a presentation in the specified order is made by the player who sends then the requests one after the other to the different multimedia servers
7. When an E_i component delivery request is received, the multimedia server delivers on line the multimedia stream from an already loaded buffer

The hosting server can use different scheduling policies to deal with the predictive request. One can consider that the more urgent request (considering the Time-begin(E_i) values) has priority. We can also state (considering Time-max(E_i) value) that among the urgent requests, the priority is given for the shortest presentations. This can be adopted, if the host server has not enough memory space to service all the requests. However, keeping the stream into memory, has a cost. To deal with that, we can establish that only presentations which need a high requirements in term of QoS, can be preloaded. In other hand, the server host can delay some predictive requests until they will reach an urging threshold. In case a server should stream the same media to different player clients, the hosting server can choose to load the longest media presentation (considering the Time-max(E_i) values) and to keep it into memory until it services all the client requests.

6 Conclusion

We have presented in this paper a formal approach for modelling a SMIL document into an Algebraic Time Net frame. The obtained ATN's frame is then translated into an equivalent time graph. The obtained graph is compact and represents all multimedia scenarios. During its construction, the graph computes for each terminal node, the minimum and maximum durations of each multimedia sub sequence. That allows us to think in either a qualitative or a quantitative analysis and scheduling of the SMIL presentation. Furthermore, the time graph could be used, at edition level in temporal formatting the multimedia document to make it consistent or at player level to produce a consistent presentation. On the other hand, some information contained in the graph could be used in the QoS predictive management.

References

1. W3C Recommendations. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification. URL:http://www.w3.org/TR/REC-smil, June 1998
2. Jourdan, M. Layaida, N. Roisin, Sabry-Ismail, L. Tardif, L. Madeus, An authoring Environment for interactive Multimedia Documents In proc of ACM Multimedia'98 Bristol, UK Sep 1998
3. D.C.A Bulterman, User-Centered Abstractions for adaptive Hypermedia Presentations In proc of ACM Multimedia'98 Brisol UK 1998 pp. 247–256
4. A.Abdelli, H. Boucheneb, and G. Berthelot, L'automate temporisé d'une spécification TC-LOTOS, in Proc. of CARI'98, INRIA ORASTOM - Dakar (Senegal) Oct 98
5. S.M. Chung and L. Pereira, Timed Petri net representation of the synchronized Multimedia Integration Language (SMIL) of XML In proc of ITCC april 2003 - USA)
6. N. Layaida, L. Sabry-Ismail, Maintaining Temporal Consistency of multimedia documents using constraint Networks In proc of the 1996 Multimedia Computing and networking, San José USA Feb 1996 pp. 124–135
7. P.N.M. Sampaio, C. Lohr, J.P. Courtiat, An integrated environment for the presentation of consistent SMIL 2.0 documents ACM Symp (DocEng'01), Atlanta(USA), Nov 2001
8. P. Merlin, D.J Farber, Recovability of communication protocols IEEE Trans on communications, 24(1976)
9. A.Abdelli, Specifying a real time system using the ATN model In Proc of the fourth international symposium on programming and systems. ISPS99 Algiers 18–21 Nov 1999
10. A.Abdelli, Constructing the contracted accessibility graph of real timed system modelled using a language based on Time Petri Nets. In Proc of I3S Fev-2001 Constantine Algeria
11. P.N.M. Sampaio, J.P. Courtiat, Providing consistent SMIL 2.0 documents 2002 IEEE (ICME'2002), Lausanne Suisse), 26-29 Août 2002, 4 p.

MULTFRC-LERD: An Improved Rate Control Scheme for Video Streaming over Wireless

Xiaolin Tong[1,2] and Qingming Huang[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences,
Beijing, P.R.China, 100080
[2] Graduate School of Chinese Academy of Sciences,
Beijing, P.R.China, 100039
{xltong, qmhuang}@jdl.ac.cn

Abstract. We propose the Loss Event Rate Discounting scheme to improve the performance of MULTFRC over wireless networks. In our MULTFRC-LERD scheme, each TFRC connection includes several loss discounting levels. When the wireless bandwidth is underutilized, we increase the discounting level of the TFRC connection at first, and when all the existing connections have reached the highest loss discounting level we open a new connection. The connection number and loss discounting level are determined by Inverse Increase Additive Decrease algorithm. Analytical and simulation results demonstrate that our scheme is a finer granularity rate control scheme for video streaming over wireless networks and can reduce the resource consumption significantly compared with MULTFRC. Moreover, LERD extends the range of applicability of the MULTFRC scheme.

1 Introduction

Over the past several years, the Internet has witnessed a tremendous growth in the use of audio and video streaming. Wide deployment of these applications without appropriate congestion control mechanisms may result in unfair bandwidth allocation or even congestion collapse. Because the congestion control scheme used by TCP halves the sending rate in response to a single loss event, it is unsuitable for the multimedia streaming applications. TCP-Friendly Rate Control (TFRC) [1], an equation-based congestion control scheme, was proposed to provide a relatively more stable sending rate for these applications.

TFRC was originally developed for media streaming over wired network. It assumed that most packet losses are due to congestion. While in a wireless environment, a significant fraction of packet losses may occur due to transmission errors. Treating wireless losses as congestive will cause TFRC reduce its sending rate unnecessarily and thus result in poor performance.

Chen and Zakhor proposed MULTFRC [2] to improve the throughput performance on the wireless link. MULTFRC allocates a certain number of TFRC connections for a given wireless streaming application to fully utilize the wireless bandwidth. Chen has shown that if the number of connections and packet

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 282–289, 2004.

size are chosen appropriately, MULTFRC can approach the optimal bandwidth utilization. The advantage of MULTFRC is that it does not need to modify the network infrastructure or protocols. However, the system resources it consumes are directly proportional to the number of TFRC connections. When the number of opened connections becomes too large, the end system, which is typically power-limited handheld equipment, can not afford such huge resource consumption. Additionally, adjusting the connection number can only achieve coarse granularity rate control, thus there is a quantization effect under the low error rate environment. Finally, the range of applicability of MULTFRC is limited.

We find that discounting the loss event rate can achieve the similar effect of opening multiple TFRC connections, and it has the potential to achieve finer granularity rate control. Furthermore, it makes MULTFRC applicable to a wider range. The cost of our scheme compared with MULTFRC is that it needs to make a small modification to the TFRC sender side code to provide the loss event rate discounting capability.

The rest of this paper is organized as follows. Section 2 reviews the related work on improving TFRC performance over wireless networks. In Section 3, we discuss the concept of Loss Event Rate Discounting (LERD) and our proposed MULTFRC-LERD scheme. Experimental results are given in Section 4. Finally, Section 5 concludes the paper and describes the future work.

2 Related Work

There have been a lot of efforts to improve the performance of TCP/TFRC over wireless networks.

Application layer solution: MULTFRC [2] can be considered as an application layer solution because it does not need to modify the network infrastructure or protocols. It improves the TFRC performance over wireless networks from the application level point of view.

Transport layer solution: In the transport layer, many efforts are devoted to differentiating the congestion losses from wireless losses. Cen et al. [3] proposed a hybrid end-to-end Loss Differentiation Algorithm (LDA), which combines three base LDAs: Biaz [4], Spike [5] and ZigZag [3]. Samaraweera [6] proposed an end-to-end noncongestion packet loss detection (NCPLD) algorithm for a TCP connection in a network with a wireless backbone link. Liu et al. [7] proposed an approach which integrates PLP and HMM. An HMM is trained over the observed RTTs to infer the cause of loss. Yang et al. [8] exploited the link layer information in wireless channels to discriminate between the wireless losses and congestion losses. Bae et al. [9] used ECN marking in conjunction with RED queue management scheme, and calculated the TCP-friendly rate based on ECN-marked packet probability instead of packet loss probability. Installing an agent at the edge of wired and wireless network is another method to discriminate the congestion losses from wireless losses [10].

3 Our MULTFRC-LERD Scheme

In this section, we first introduce the concept of Loss Event Rate Discounting (LERD), and then present the framework of MULTFRC-LERD system. Finally, we describe the IIAD control algorithm used in our system.

3.1 LERD: Loss Event Rate Discounting

We use the following simple TFRC model to explain the concept of LERD:

$$T = \frac{c \cdot S}{rtt \cdot \sqrt{p}} , \tag{1}$$

where p denotes the Loss Event Rate, T represents the transmit rate, S is the packet size, rtt is the end-to-end round trip time, and c is a constant factor. This simple model has captured all the essential factors and can explain the LERD scheme more clearly.

When the wireless link is underutilized, the total throughput of multiple TFRC connections is multiplying the throughput of one TFRC connection by the number of connections m. Thus the throughput of MULTFRC can be expressed as:

$$T_m = m \cdot \frac{c \cdot S}{rtt \cdot \sqrt{p}} . \tag{2}$$

By equation analysis and simulation validation, we find that the similar effect of multiple simultaneous TFRC connections can be achieved by discounting the loss event rate p, i.e. ,

$$p_d = d \cdot p , \tag{3}$$

where d is the discounting factor, and it can take any value between 0 and 1. In the case of original TFRC, the discounting factor d is equal to 1. The smaller discounting factor corresponds to the more aggressive TFRC-LERD connection. Replacing the loss event rate p in Equation 1 by p_d, we get the throughput of one TFRC-LERD connection:

$$T_d = \frac{c \cdot S}{rtt \cdot \sqrt{p_d}} . \tag{4}$$

Comparing Equation 2 with Equation 4, we can see that MULTFRC and LERD try to improve the throughput by adjusting different parameters. MULTFRC uses the number of connections m to tune the throughput, and our LERD scheme uses the discounting factor d to adjust the sending rate. The discounting factor d can take any value between 0 and 1.0, however, the number of connections m in MULTFRC scheme can only take the positive integers, e.g., 1, 2, 3,$\cdots$, therefore LERD has the potential to achieve finer granularity rate control than MULTFRC.

3.2 MULTFRC-LERD

The framework of MULTFRC-LERD system is illustrated in Fig. 1. In our system, each TFRC connection is divided into L loss discounting levels. Each level is characterized by a discounting factor $d_i (1 \leq i \leq L, 1 = d_1 > d_2 > \cdots > d_L > 0)$. The lowest level is assigned $d_1 = 1$, which means no discounting at all. For each TFRC connection C_k we maintain a variable d_{c_k} in its sender side to keep its current loss discounting factor. When TFRC sender receives the loss event rate p from the receiver, we discount it by its current discounting factor, i.e., $p_d = p \cdot d_{c_k}$, and use the discounted loss event rate to compute the sending rate. The number of connections and their respective discounting factors are controlled by the Connection Manager.

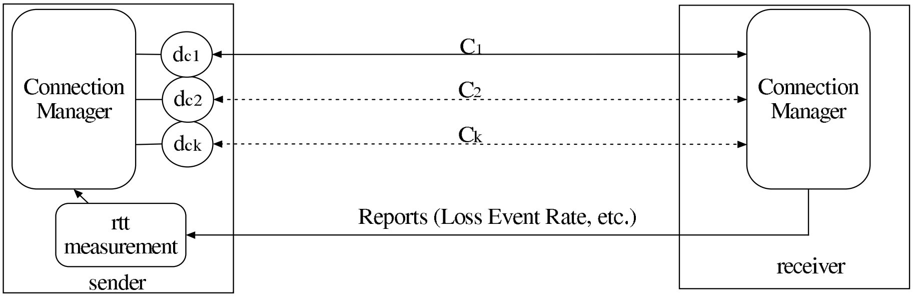


Fig. 1. The framework of MULTFRC-LERD system

The control algorithm employed in our Connection Manager is the Inverse Increase Additive Decrease (IIAD) algorithm similar to that used in MULTFRC. But the adjusting granularity in our algorithm is one discounting level rather than one connection as in MULTFRC. Our IIAD algorithm is described as follows (*threshold*, *alpha* and *beta* are predetermined constants.):

```
Measure average_rtt over a specified period.
if (average_rtt < threshold) {
  discount_level = discount_level + alpha / discount_level;
}
else {
  discount_level = discount_level - beta;
}
```

When we need to increase the discounting level, which is indicated by the IIAD algorithm, we increase the discounting level of the last opened connection. If the last opened connection has reached the highest discounting level, we open a new TFRC connection. Similarly, when we need to decrease the discounting level, we decrease the discounting level of the last opened connection. If the last opened connection has reached its lowest discounting level, we close this connection.

4 Simulation Results

In this section, we implement the loss event rate discounting functionality and compare the performance of MULTFRC and MULTFRC-LERD.

4.1 Simulation Setup

We use ns-2 to study the performance of MULTFRC-LERD. The network topology used in the simulation is shown in Fig. 2, which is the same as that used in MULTFRC. The transmission error in the wireless link is simulated by the exponential error model, and the packet error rate varies from 0.01 to 0.16. The last hop wireless link is the bottleneck of the path from sender to receiver. The parameters used are shown in the figure. In this simulation, the values of *threshold*, *alpha* and *beta* are $1.2 \cdot rtt_{min}$, 1, and 1, respectively. We employ a simple three-level discounting scheme, and the corresponding discounting factors are 1.0, 0.5 and 0.2. It is possible to use more refined discounting levels. Here we just want to demonstrate the effectiveness of LERD.

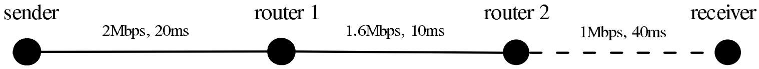


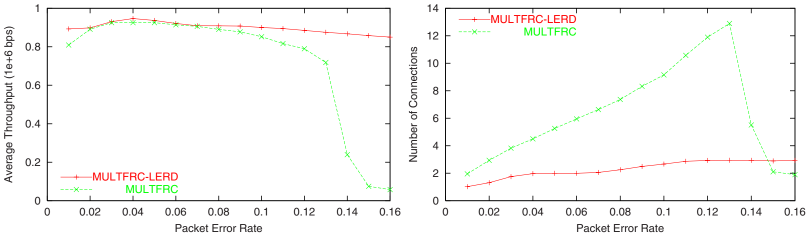
Fig. 2. Wireless last hop topology

4.2 Performance of MULTFRC-LERD

We simulate MULTFRC-LERD and MULTFRC for 9000 seconds with the packet error rate varying from 0.01 to 0.16. The performance comparison is given in Fig. 3.

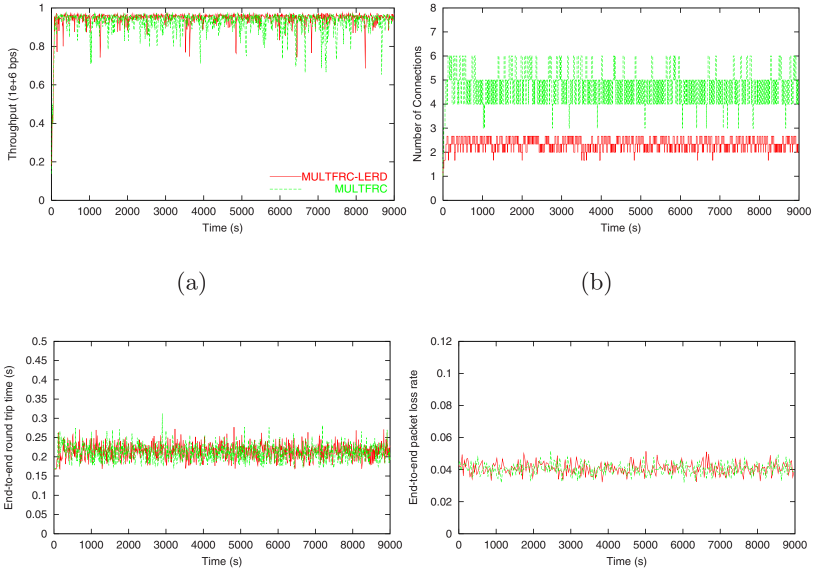
Fig. 3(a) shows the average throughput of our MULTFRC-LERD scheme and MULTFRC. Notice that when the packet error rate is low our scheme outperforms MULTFRC, because the connection number can only achieve coarse granularity rate control and there is a quantization effect when the number of connections is small, which is also pointed out in [2]. The throughput improvement of our scheme is achieved by its finer granularity rate control. We find that these two schemes have comparable performance when the wireless packet error rate is between 0.02 and 0.08. But when the error rate exceeds 0.08, the performance of MULTFRC degrades. It is because the original TFRC cannot work well under such high packet error rate environment, where the high loss event rate prevents TFRC from increasing its sending rate. However, LERD scheme mitigates the negative impact of high wireless error rate by discounting the loss event rate and is still able to approach the optimal bandwidth utilization even when the error rate is very high.

Fig. 3(b) illustrates that MULTFRC-LERD uses much fewer connections than MULTFRC. When the packet error rate exceeds 0.13, the connection number of MULTFRC drops sharply, because under such high error rate the MULTFRC scheme is not effective any more, its connection number cannot reach the optimal value and the performance degrades significantly.



(a) (b)

Fig. 3. MULTFRC-LERD and MULTFRC under different packet error rate environment (from 0.01 to 0.16): (a) average throughput. (b) average number of connections opened



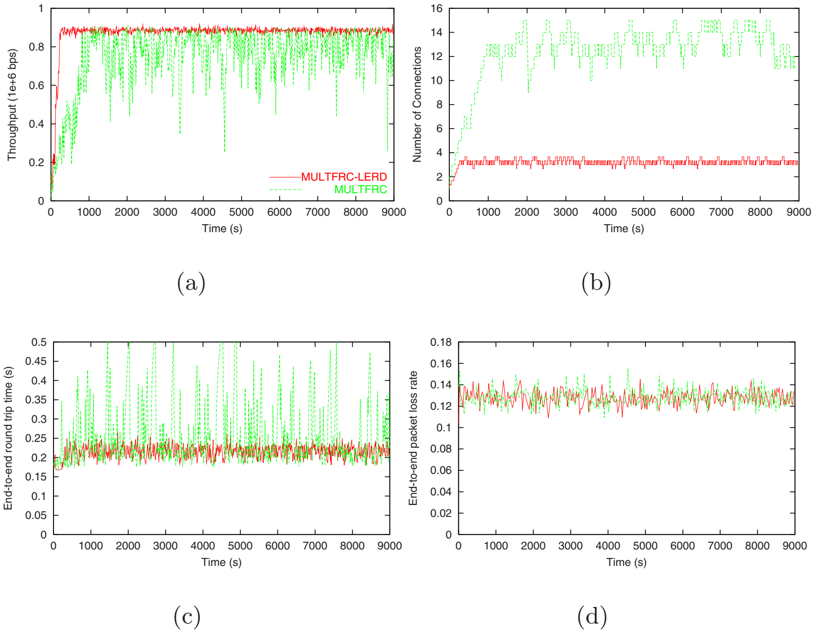
(a) (b)

(c) (d)

Fig. 4. The performance comparison of MULTFRC-LERD and MULTFRC when the wireless packet error rate is 4%: (a) throughput, (b) number of connections opened, (c) end-to-end round trip time, (d) end-to-end packet loss rate

In the following, we analyze two typical scenarios: the medium error rate environment (PER=0.04) and the environment with very high error rate (PER=0.12).

Fig. 4 demonstrates that the performance of our MULTFRC-LERD scheme is comparable to MULTFRC under the error rate of 0.04, but the number of connections opened by MULTFRC-LERD is much smaller. Note that in Fig. 4(b),



(a)

(b)

(c)

(d)

Fig. 5. The performance comparison of MULTFRC-LERD and MULTFRC when the wireless packet error rate is 12%: (a) throughput, (b) number of connections opened, (c) end-to-end round trip time, (d) end-to-end packet loss rate

we also display the discounting levels of MULTFRC-LERD. Because the optimal point of MULTFRC resides between 4 and 5, MULTFRC frequently adds and drops connection to maintain a balance. However, in our scheme most adjustment is accomplished by changing the discounting factor. We argue that it is favorable because the cost of opening and dropping a connection is much higher than setting the discounting factor to a new value.

Fig. 5 illustrates that the performance of MULTFRC-LERD is much better than MULTFRC when the error rate is as high as 0.12. Under such environment, the original TFRC does not function properly any more. The end-to-end round trip time varies greatly, the connection number changes irregularly, which in turn results in highly varying throughput. However, our scheme is quite stable under this abnormally high error rate environment because LERD enhances the capability of TFRC to deal with high error rate.

5 Conclusion and Future Work

In this paper, we have proposed the Loss Event Rate Discounting scheme to improve the performance of MULTFRC system. We tested our scheme extensively under different error rate environments. Our results demonstrate that the advantages of our MULTFRC-LERD scheme over MULTFRC are three-folded:

Firstly, it reduces the resource consumption. The resource consumed by MULTFRC is reduced by 50 to 80 percent. Secondly, it can achieve finer granularity rate control. Our scheme mitigates the quantization effect of MULTFRC and improves its throughput under the low error rate environment. Finally, our scheme works well under very high error environments where MULTFRC does not function properly any more. In future work, we will investigate the more refined discounting levels and discounting factors, and how they affect the performance of MULTFRC-LERD.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China under grant No. 60333020, the National Fundamental Research and Development 973 Program of China under grant No. 2001CCA03300, and the Natural Science Foundation of Beijing under grant No. 4041003. The first author would like to thank Minghua Chen for his helpful discussion.

References

1. S. Floyd, M. Handley, J. Padhye, and J. Widmer, Equation-Based Congestion Control for Unicast Applications, In Proceedings of ACM SIGCOMM, August 2000
2. M. Chen and A. Zakhor, Rate Control for Streaming Video over Wireless, Proceeding of Infocom 2004, March 2004
3. S. Cen, P. C. Cosman, and G. M. Voelker, End-to-end differentiation of congestion and wireless losses, IEEE/ACM Transactions on Networking, Volume 11, pp. 703-717, October 2003
4. S. Biaz, N. H. Vaidya, Discriminating congestion losses from wireless losses using inter-arrival times at the receiver, Proc. of IEEE Symposium on Application-specific System and Software Engr. and Techn., pp. 10-17, Mar 1999
5. Y. Tobe, Y. Tamura, A. Molano, S. Ghosh, and H. Tokuda, Achieving moderate fairness for UDP flows by path-status classification, in Proc. 25th Annual IEEE Conf. on Local Com-puter Networks, pp. 252-261, Tampa, FL, Nov 2000
6. N. Samaraweera, Non-congestion packet loss detection for TCP error recovery using wire-less links, in IEE Proceedings of Communications, 146(4), pp. 222-230, Aug 1999
7. J. Liu, I. Matta, and M. Crovella, End-to-end inference of loss nature in a hybrid wired/wireless environment, Boston Univ., Boston, MA, Comput. Sci. Dept., Tech. Rep., Mar. 2002
8. F. Yang, Q. Zhang, W. Zhu and Y.-Q. Zhang, Bit Allocation for Scalable Video Streaming over Mobile Wireless Internet, Proceeding of Infocom 2004, March 2004
9. S. J. Bae and S. Chong, TCP-Friendly Flow Control of Wireless Multimedia Using ECN Marking, Signal Processing: Image Communication, Vol. 19, No. 5, pp. 405-419, May 2004.
10. G. Cheung and T. Yoshimura, Streaming agent: A network proxy for media streaming in 3G wireless networks, in Proc. IEEE Packet Video Workshop, Pittsburgh, April 2002

Multi-source Media Streaming for the Contents Distribution in a P2P Network

Sung Yong Lee, Jae Gil Lee, and Chang Yeol Choi

Kangwon National University
Department of Computer, Information and Telecomm. Engineering
192-1 Hyoja2-dong, Chuncheon 200-701, Republic of Korea
moota4@mmslab.kangwon.ac.kr

Abstract. For a P2P network, the contents distribution is a very important service because the contents provider is not fixed. And media streaming and file saving operations should be simultaneously carried out in order to distribute contents through the P2P network. In this paper, a P2P multi-source media streaming system which can replay the contents data during downloading is proposed and implemented. The system reduces the user response time and the number of simultaneous user increases more than two times.

1 Introduction

As P2P(Peer-to-Peer) based contents sharing programs are actively used on the internet, interest on the P2P network is being much more increased recently. And each peer has a duty to perform roles as a server as well as a client, therefore each peer takes the contents from the other peer and gives the contents to the other peer. At that time if the media data is received by only a specific peer user QoS is not satisfactory and a one-to-one P2P paradigm can't guarantee data rate and reliability needed by user because of frequent network escape of peer. To solve these problems multi-source downloads or multi-source streaming to receive data from several source peers are proposed [1] [2]. eDonkey [5] downloads data from several source peers, makes a temporary file, and enhances the speed by merge of temporary files into a file after receiving all data. GNUSTREAM [3] receives different data segments from each of source peers, then assembles and replays it. Like this, eDonkey and GNUSTREAM enhance transmission rate and mitigate the damage of the source peers escape by receiving data from multi-sources. But in the case of eDonkey, replay should be performed after downloading all the files, therefore response time becomes long. On the other hand, GNUSTREAM can't distribute contents on the P2P network because it is hard to share contents with other peers as it consumes media data after replay like a server-client streaming.

On the P2P network the contents distribution is a very important service because the contents provider is not fixed unlike CDN(Contents Delivery Network). And in the P2P media streaming, a request peer replays and saves data simultaneously, and after streaming it acts as a new source peer providing files

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 290–297, 2004.

to other peers. Therefore media streaming and file saving operations should be simultaneously carried out in order to distribute contents through the P2P network. If the number of peers which possess a specific file is small, the number of peers that participate in multi-source streaming is also small, so the number of peers that receive and provide specific data may be limited. As the existing contents sharing models can replay after downloading the data, the response time becomes long and the contents distribution is difficult.

In this paper, a P2P multi-source media streaming system which can replay contents during downloading is proposed and implemented. In the proposed system, as the request peer receive, replay and save the data files, it can play a role as a source peer as well as streaming. As a result, user response time can be reduced and fast contents distribution through network is possible by a mechanism transmitting only a part of media files.

2 Design and Implementation

2.1 Basic Operations

Every peer performs transmitting and receiving data simultaneously. The receiving peer requests a data block that divided media files in small size to a source peer, receives the other parts of media files from other peers in parallel, and replays it. It is possible to transmit the media when source peers have all parts of media file as well as when they have a part of a media file.

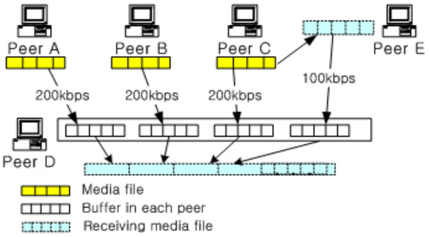
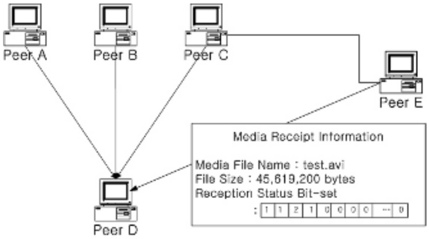


Fig. 1. Media data flow among peers



Media Receipt Information

Media File Name : test.avi
File Size : 45,619,200 bytes
Reception Status Bit-set
: 1 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | ··· | 0

Fig. 2. Exchanges for the media receiving states information

In the Fig. 1 Peer D is a receiving peer and plays as a client, Peer A, B, C are source peers and transmit a part of media file assigned to Peer D. Peer E receives data from peer C, stores and transmits the data to Peer D. Peer D writes data received from Peer A, B, C in the buffer block for each peer, stores it as a file, and replays the file. Peer E performs transmission and receipt simultaneously, and Peer D receives media-receipt-information from Peer E as shown in Fig. 2. And it confirms the states of media store, and by comparing it with media state information bit set, Peer D determines whether to request the next media or not.

2.2 Peer to Peer Communication

- Media information request : The requested file name is transmitted to the selected source peers.

```
j = 1 to N(Selected Source Peers's Number)
  if (Connection State == TRUE) then
    Message = PacketSize + Type(File_Info_Req) + FileName
Send message to selected peers that are connected
```

- Media information response : If there is an intact file, whether transmission is possible or not and file information are transmitted. For the receiving file, the file information and receipt state bit set(BitStream_Index) are together transmitted.

```
if (State == SEND_OPEN Number)
  if (State of Media file == Entire File) then
    Message = PacketSize+Type(File_Info_Ack)+State of File(Entire File)
           +File Name+FileSize
  else if (State of Media file != Entire File) then
    Message = PacketSize+Type(File_Info_Ack)+State of File(Not entire File)
           +File Name+FileSize+BitStreamIndex
Send message to request peer
```

- Media data request : A part of the media file to be transmitted is assigned to each peer and transmitted with the file name. After checking the receipt state bit set during the media receiving, it assigns and transmits the index of media file whose bit is '0' to the source peer.

```
case state bit is 0: not yet requested
case state bit is 1: corresponding part has been completely received
case state bit is 2: corresponding part is being now received
for j=1 to N(Selected Source Peers's Num.)
  Message = PacketSize+Type(File_Stream_Req)+FileName(Request FileName)
         +Start Pointer+End Pointer
Request again after checking my Bit_Stream_Index
```

- Media data transmission : After the separating of the Start Index and the End Index from the media stream request packet, the assigned data is copied from a memory map file. The Start Index and the End Index are kept in the source peer until the transmission of the assigned parts completes. Here content size means the size of media data transmitted in a packet.

```
Message = PacketSize+Type(FILE_STREAM_ACK)+Send Point+Offset
       +Content Size+Media Data
Send message to request peer
```

- Media receipt state information update request : The request peer compares its own media receipt state bit set with the source peers, and it examines the block to request to the source peer. If there is no more block to request, it requests the media receipt state information update.

```
Message = PacketSize + Type(Update_bitStream_index_req)+ FileName
Send message to request peer
```

– Media receipt state information update transmission : If there is a request of the media receipt state information update, it converts the current media receipt state bit set into '0's and '1's and transmits it to the request peer.

```
Message = PacketSize+Type(Update_bitStream_index_Ack)+FileName+Bit_Stream_index
Send message to request peer
```

2.3 Client Module

The proposed system was implemented with Visual C++ and C on Windows XP. All the peers have the server module and the client module, the server module is a routine to perform commands requested by other peers and the client module performs the commands of source peers.

Selection of the source peers. To select the source peer it is examined whether the file can be transmitted or not immediately and whether the number of source peers exceeds the number of receipt peers connected by source peers in maximum or not. The network bandwidth and proximity with receipt peer are also examined. Available network bandwidth can be calculated with a TCP-friendly rate control algorithm(TFRC) [3].

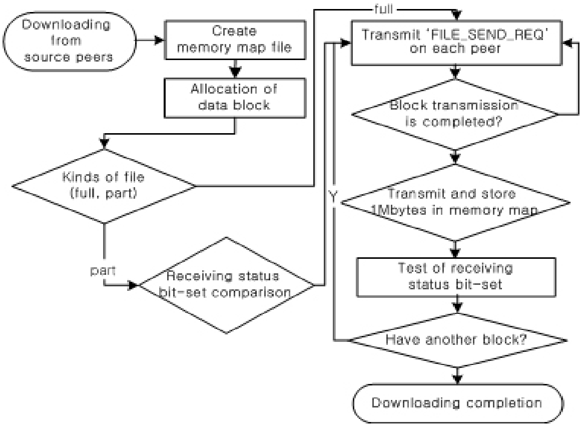

Downloading from source peers
Create memory map file
Allocation of data block
Kinds of file (full, part)
full
part
Receiving status bit-set comparison
Y
Transmit 'FILE_SEND_REQ' on each peer
Block transmission is completed?
Transmit and store 1Mbytes in memory map
Test of receiving status bit-set
Have another block?
Downloading completion

Fig. 3. Assignment of data block

Assignment of data block for the transmission. Figure 3 shows the procedure for assigning a data block to each source peer. First all, constants block size (1Mbytes) is assigned to each source peer. For source peers whose transmission finished early, media receipt state bit set is examined, and the earliest block among ‘0’ areas is assigned. The next data block is assigned to the fastest source peer.

The assignment of data blocks can be categorized into two fields as the file types. For the *mpg* file the initial part is assigned by 1Mbytes one after another. On the other hand in the case of the *avi* file which obeys the RIFF file standard, video/audio part is assigned by 1Mbytes one after another after receiving the header and *avi* index. An analysis [4] of file types requested by users in Gnutella said that more than 65% of the users have requested multimedia contents and among the contents requested *avi* files was the most frequent as 18.72%. Our system can also support *asf* type files as well as *avi*, *mp3*, and *mpg* types which are the most frequently requested.

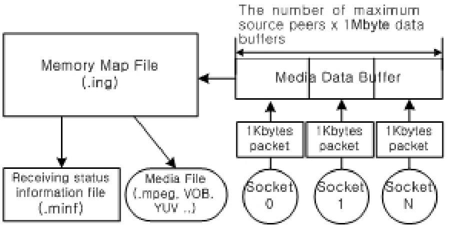

The number of maximum source peers x 1Mbyte data buffers
Memory Map File (.ing)
Media Data Buffer
1Kbytes packet
1Kbytes packet
1Kbytes packet
Receiving status information file (.minf)
Media File (.mpeg, VOB, YUV ..)
Socket 0
Socket 1
Socket N

Fig. 4. Media data merge

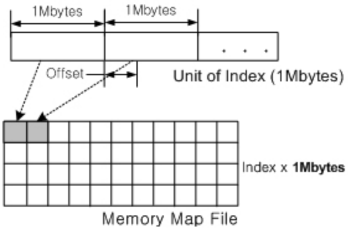


Fig. 5. Data transmission between the buffer and the memory map file

Data merge and storing the file. Every peer has a data buffer and a memory map file as shown in Fig. 4. The data merge and the file store are processes of assembling the media data from the source peers one after another, and the data transmitted by client socket is written in the data buffer block assigned to each source peer. When the data buffer is full, the buffer data is written in the corresponding location of memory map file. After a while the data is replayed by opening the memory map file. The data between the data buffer and memory map file is moved as shown in Fig. 5.

2.4 Server Module

The server module is composed of the file information transmission, peer connection, the data block assignment, and data transmission. Transmission of the file information is a process to know the identity of the file, whether the transmission is possible or not and the peer state after the request peer searches media, which can be divided into the information transmission of the downloaded file that is an intact and the information transmission of the downloading file itself. In the peer connection step, it should not exceed the maximum number of transmission peers. In the data transmission step, the transmission rate is limited by delay time between packets in order to protect the resources of the source peer. After calculating the file seek pointer in 1Mbytes data block index the corresponding 1Mbytes file is transmitted in 1Kbytes packets after delay time.

3 Experiments and the Results

A simple P2P network for experiments was organized as shown in Fig. 6. Network A and B are connected as to T3 class and the inside of network A and B is connected to 100Mbps Ethernet. The peers of each network are composed of Pentium IV, with more than 256MBytes DRAM and Windows XP.

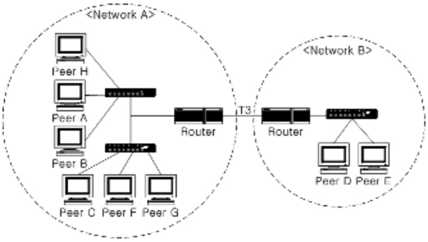


Fig. 6. Configuration of the testbed system

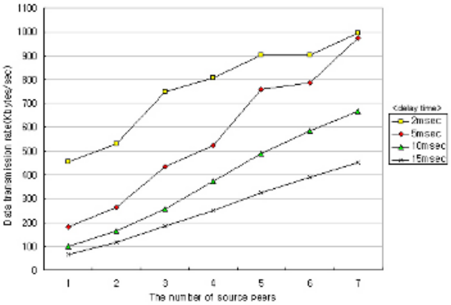


Fig. 7. Changes of the data transmission rate

3.1 Data Transmission Rates

The data transmission rates are measured as a time that receipt peer writes 30MBytes data in the file, increasing the number of source peers from 1 to 7, changing the delay time between data packets to 2, 5, 10, 15msec. As shown in Fig. 7, when the delay time is 2 msec, if increases the number of source peer from 1 to 7, the data transmission rate increased by 683%.

3.2 Media Streaming

The minimum number of the source peers and the delay time between packets of the source peer which enable files encoded with MPEG–1, MPEG–2, DivX to download and replay simultaneously are found. The relations between time variations needed to download all files and the number of source peers are observed. The results are shown in Table 1.

3.3 The Simultaneous Download and Upload

Fig. 8 shows an example system which several users can replay the media file simultaneously. The test file is amuro.mpg which size is 352 × 240, and the frame rates are 30fps, bit rates are 172KBytes/s. Even though there is only one media file originally, 4 peers could replay media files simultaneously through simultaneous download and upload. This feature increased the distribution speed of contents and the number of simultaneous users more than two times. Fig. 9 and Fig. 10 show the client and server view of intermediate node respectively, and Fig. 11 shows a client view of the bottom node.

Table 1. Characteristics of download and simultaneous replay for MPEG–1,2, DivX and Avi

File name	Number of source peers	Delay time between packets	Data transmission rate(Kbytes/s)	Download time(sec)	File replay time(sec)
Amuro.mpg (MPEG–1, Low Resolution)	1	5ms	182.85	250	268
	2	5ms	262.56	174	
	3	15ms	187.88	243	
Boa.mpg (MPEG–1, High Resolution)	1	2ms	455.11	147	259
	2	5ms	262.56	255	
	3	10ms	259.24	258	
Sheri.VOB	3	2ms	749.26	392	407
	4	2ms	808.42	364	
Ani.avi (Divx4, mp3)	1	15ms	66.06	387	1203
X-file (Divx3, mp3)	1	5ms	182.00	184	314
	2	15ms	117.70	284	
badboy.avi (Xvid, mp3)	1	5ms	192.00	520	622
	2	10ms	190.00	526	
	3	15ms	208.00	480	
Firstlove.avi (Divx3, AC3)	1	2ms	455.11	340	629
	2	5ms	262.56	589	
	3	10ms	259.24	597	

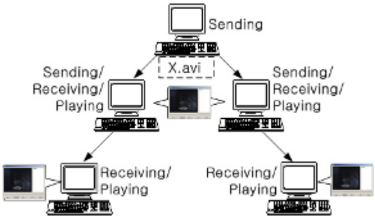


Fig. 8. System configuration

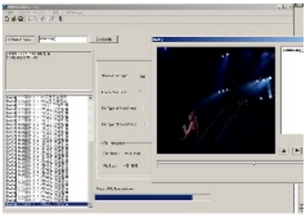


Fig. 9. Client view of an intermediate node

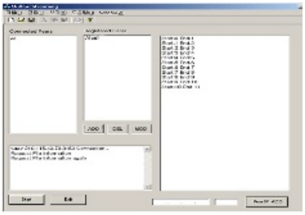


Fig. 10. Server view of an intermediate node

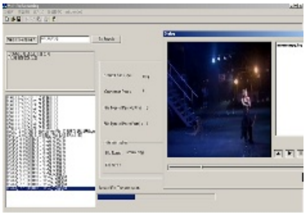


Fig. 11. View of a bottom client

4 Conclusions

The existing contents sharing programs had a shortcoming that the response time is long and it is not possible to confirm whether the contents is what the user wants or not during download. On the other hand, the existing researches on P2P multi-source media streaming did not considered the contents distribution services.

The proposed system can perform data transmission and receipt, file store and file replay simultaneously, therefore it can distribute contents between P2P peers easily. The system has an advantage to reduce the user response time and several features as it follows. First, it provides higher media transmission rates than that needed for file replay because it receives data from several peers. Experimental results show that the data transmission rate increased more than 683%. Second, the media can be replayed in real-time by saving received media as a file simultaneously. In addition, the contents distribution is also possible. Third, transmitting a part of media file which downloading peer has, makes fast distribution and diffusion of contents possible and the number of simultaneous user increases more than two times.

Acknowledgments. This work was supported by the Ministry of Information and Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.

References

1. D.Xu, M, Hefeeda, S.Hambrusch and B.Bhargava, “On Peer-to-peer Media Streaming”, *IEEE ICDCS 2002*, pp. 363–371, July 2002
2. Reza Rejaie, Antonio Ortega, “PALS:Peer-to-peer Adaptive Layered Streaming”, *Proceedings of the International Workshop on Network and Operation Systems Support for Digital Audio and Video*, June 2003
3. T. Nguyen and A. Zakhor, “Distributed Video Streaming Over Internet”, *Proceedings of SPIE Conference on Multimedia Computing and Networking*, January 2002
4. Demertis Zeinalipour-Yazti, Theodoros Folias, “A Quantitative Analysis of the Gnutella Network Traffic”, CS204 final project, California University, USA, July 2002
5. eDonkey, “http://www.edonkey2000.com”

A New Feature for TV Programs: Viewer Participation Through Videoconferencing

Jukka Rauhala and Petri Vuorimaa

Telecommunications Software and Multimedia Laboratory
Helsinki University of Technology,
P.O. Box 5400, FI-02015 HUT, Finland

Abstract. In this paper, the inclusion of a videoconferencing server is proposed as a new feature for TV broadcast systems. With this feature, TV broadcast stations can enable viewers to participate in TV programs by using a videoconferencing client. Some program concept ideas, such as game, talk, and radio shows, are analyzed in this paper from the feature's point of view. The technical implementation of the feature is studied by sketching the implementation in a real broadcast system as well as creating a reference implementation, which is tested in a real TV broadcast system. The results show that the feature is possible to implement and it suits well into TV environment, because it provides participation with visual aspect. Hence, it can be easily seen that this feature has its place in a future TV broadcast system.

Keywords: Videoconferencing, TV broadcast, SIP

1 Introduction

Many TV programs enable viewer participation. So far, the three alternatives for participation have been phone call, Short Message Service (SMS), and moving camera. In the first one, the viewers call a certain phone number to get their voice into the live TV broadcast. In the second one, the viewers send a SMS, which is shown in the TV broadcast. In the third alternative, a reporter interviews people in a live TV broadcast with a moving camera. The reason why this feature is popular is that it offers ordinary viewers a chance to become famous for a short moment; a chance to be seen or heard in a live TV broadcast.

The key idea of this paper is to extend viewer's participation through videoconferencing. It has several advantages over other alternatives. First, it enables participation with a plain videoconferencing device regardless of location, which means that in the future it will be possible for everyone. Second, it provides not only live audio, as with a normal voice connection, but also live video. Third, it provides more rich content for the TV broadcast with a low cost for the station. Fourth, it enables to create new kind of concepts for TV programs.

This feature needs at the simplest only two extra elements for the TV broadcast system: videoconferencing server and a video mixer. Furthermore, the only requirement for the videoconferencing client is interoperability with the server

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 298–305, 2004.

in the broadcast system. Hence, the merging of videoconference into a broadcast system has been technically possible for some time.

The feature introduced in this paper will work in both analog and digital TV broadcast systems, but since television is evolving to its digital counterpart, the focus is on the digital TV. In addition, digital TV includes interactive applications, which add value to this feature. For example, the viewer could get additional information about the person on the screen by clicking a certain button.

Videoconferencing is a widely researched topic. International Telecommunications Union (ITU) began to work with their first videoconferencing protocol, H.320, in 1990 followed by Internet Engineering Task Force (IETF) a couple of years later [1],[2]. On the other hand, broadcast with participation has not been a popular subject for research, though some research has been done on gaming, an example is the virtual game show developed by Greenhalgh et al. [3].

2 Background

2.1 Digital TV Broadcasting

The main purpose for the TV broadcast system is to transmit the source audio-visual (A/V) signal through a network, such as terrestrial, cable, or satellite. In digital TV, the signal is broadcasted digitally, and the broadcast stream includes interactive applications as well. In addition, the digital TV broadcast requires less bandwidth, because the transmitted signal is compressed.

Figure 1 shows Otadigi [4],[5] as an example of a digital TV broadcast system. It is a Digital Video Broadcasting - Terrestrial (DVB-T) compliant academic and research system located in the Helsinki University of Technology campus area. In Otadigi, the source A/V signal is first encoded into MPEG-2 format using the MPEG-2 encoder. Then, the application stream from the Object carousel is combined with the encoded stream in the Multiplexer. Finally, the combined

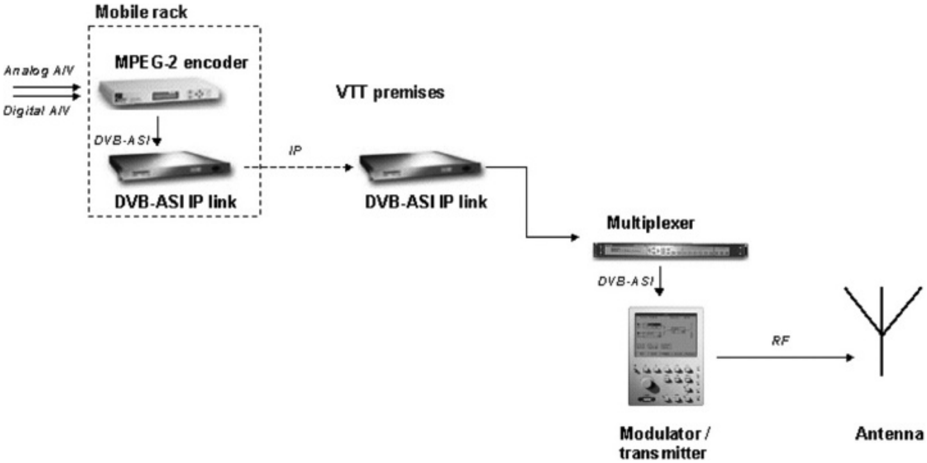


Fig. 1. Otadigi Digital TV broadcast system

stream is modulated with the Modulator and transmitted via antenna. Otadigi uses DVB-ASI IP links to transmit data between different physical locations.

2.2 Videoconferencing

Videoconferencing is an audiovisual communication service between two or more parties. While videotelephony is the essential part of the service, it can provide also additional services, such as file sharing. Videoconferencing systems can be divided into two groups: dedicated group videoconferencing systems [1] and personal videoconferencing systems. This paper focuses on the latter group, which includes videoconferencing using Personal Computers (PC), mobile phones, and digital TV receivers.

Videoconferencing service needs three elements: signaling protocol, video codec, and audio codec. The most common signaling protocols are Session Initiation Protocol (SIP) [6] and H.323 [7]. H.263 and MPEG-4 are examples of popular video codecs. G.723 and G.711 are well known audio codecs. In order to provide the service, a videoconferencing device needs video camera, microphone, loudspeaker, screen, and network connection.

Personal videoconferencing has not yet become as popular as some people have predicted [8]. The reason for this is the lack of popular commercial devices providing videoconferencing. PCs have this feature, but it has not become successful, because it requires technical skills in PC environment. However, by including this service in future mobile phones and digital TV receivers, the number of videoconferencing users will increase rapidly.

2.3 Videoconferencing Signaling Protocols

The two major protocols for personal videoconferencing, SIP and H.323, are coming from two different telecommunication fields. The former was originated in the Internet world, while H.323's background is in telecommunication networks. Because of their different origins, both protocols have different strengths and weaknesses. These protocols are widely used at the moment, and it seems unlikely that either one of them will become dominant in the near future.

SIP is a control and signaling protocol developed by IETF. The main purpose of SIP is to create, modify and terminate sessions with one of more participants. Its main advantages include scalability, extensibility, and flexibility. In addition, it is an open standard very similar to HyperText Transfer Protocol (HTTP). Currently, the main problem of SIP is the lack of interoperability support with Public Switched Telecom Network (PSTN). [6]

H.323 is an ITU recommendation for multimedia communication signaling [7]. It relies heavily on several other ITU recommendations, such as H.245, H.225.0, H.332, H.450.1, H.450.2, H.450.3, H.235, and H.246. The signaling and the call setup are based on Q.931 Integrated Services Digital Network (ISDN) signaling protocol. The strengths of H.323 are full backwards compatibility with older versions, and interoperability with PSTN [9]. The main weakness of H.323 is its complexity.

3 TV Program Concepts

Many TV programs provide participation through phone line. Videoconferencing can replace phone line in all cases, and it is also more suitable for TV environment. Another common way to enable participation is to have a moving camera with a reporter, which makes it very limited as it needs expensive equipment, and it is tied to a certain location. The main benefits of videoconferencing over moving camera are the cheap price for the broadcast company, and that it is accessible for a large audience regardless of their location. However, its main drawback is the audio and video quality compared to the moving camera.

3.1 Game Shows

In the current TV game shows, viewers can participate through a phone line. The viewer plays the game with the phone as a controller and the game screen is broadcasted live. In addition, the host of the game is able to interview and hear the viewer's comments through the phone line.

Using the technology proposed in this paper, this concept can be extended to use videoconferencing instead of phone line. This extended version enhances the viewer's participation level and it is more suitable for TV. It would make the game also more interesting, as the live video of the player would be added into the game screen, and the player would create more drama by his actions and expressions. In addition, it enables the creation of new game show concepts, which take advantage of the videoconferencing. For example, the game could require some kind of visual response, such as acting, from the player.

3.2 Talk Shows

Different kinds of talk show concepts use phone line participation. For example, in political talk show related to upcoming elections, the viewers are able to give comments or ask questions. Again, by using videoconferencing instead of phone line this concept would suit better into TV. Moreover, it would make possible to create programs, which could rely more in the interaction with the viewers. For example, a program could simply have a politician answering viewers' questions.

3.3 Radio Show Concepts in TV

New kind of concepts in TV could be obtained by transferring ideas, which are originally radio shows using the participation via phone line. An example of this is a Finnish radio program “nature night”. It is a program where specialists answer people's questions related to nature. This concept has been brought to TV already, but the TV version did not bring much more value since the interaction was still through the phone line.

With videoconferencing, the TV version could work much better as the people asking questions are being broadcasted as well. Moreover, the people would be able to show pictures, for example of animals, in live TV broadcast.

4 Technical Implementation

4.1 Requirements

In order to provide the new feature introduced in this paper, a new subsystem is needed in the broadcast system. The subsystem, shown in Figure 2, includes two elements: videoconferencing server and video mixer. The videoconferencing server handles the signaling for the videoconferencing, receives the encoded A/V signal from the client, and decodes it into raw format. Then, the decoded A/V signal is mixed with the broadcast A/V stream in the video mixer into one stream, which can be broadcasted. In addition, the video mixer will decide the location and the size of the videoconferencing stream.

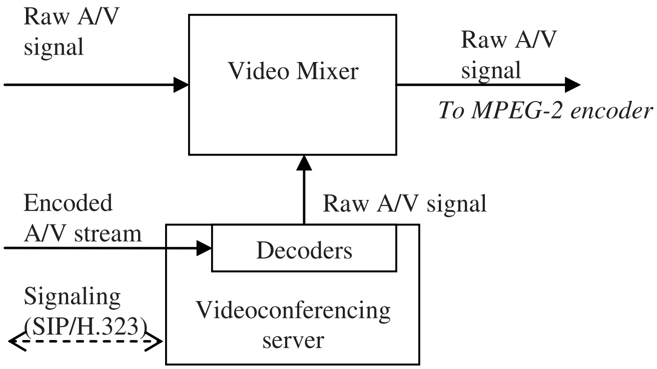


Fig. 2. Technical implementation for merging videoconferencing into broadcast stream

The server needs to support at least either one of the most common videoconferencing protocols (i.e., SIP and H.323). The best solution would be to support both protocols. In addition, the server needs to support as many of the common audio and video codecs as possible.

In the digital TV broadcast system, the best location for this would be before MPEG-2 encoding takes place. In the broadcast system reference model shown in Figure 1, it would be between the A/V signal and the MPEG-2 encoder.

4.2 Implementation in a Commercial Broadcast System

When the implementation of this feature in a commercial broadcast system is considered, a list of requirements can be identified. First, the station wants to control the access to the videoconferencing server. This would be done by receiving the videoconferencing calls in a call center, and then redirecting the accepted calls to the videoconferencing server. Second, the quality of video and audio quality signal should be very high. This means that professional hardware video mixer and decoders are used. Third, the TV program director would like to have control over the location and the size of the videoconferencing video stream on the screen. This is possible by using the hardware video mixer. Fourth, the system must be interoperable with most available videoconferencing clients. Hence, it must support both SIP and H.323 as well as all common audio and video codecs. Finally, the one-way delay from the client to the TV broadcast should be 150 ms at the maximum [10].

4.3 Reference Implementation

A reference implementation was implemented in order to show the feasibility of the ideas proposed in this paper. The implementation is shown in Figure 3. The video mixer and the videoconferencing server were included in the same software, which is running in a Linux PC. The software was written in Java and it uses Java Media Framework (JMF) for audio, video, and Realtime Transport Protocol (RTP) handling.

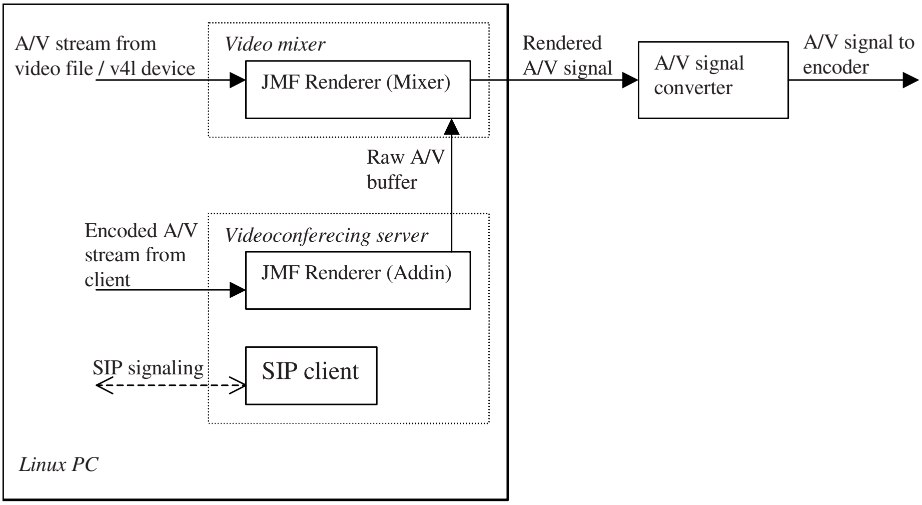


Fig. 3. Reference implementation

The signaling in the videoconferencing server is handled by a simple SIP client, which is based on previous work done in the same project [11]. When it gets SIP INVITE message, it will automatically accept the incoming call and start receiving audio and video streams. After the streams are received, JMF will set up a path, where the data is first decoded, and then passed to the customized renderer.

The mixer accepts broadcast stream input from a video file recognized by JMF or from Video for Linux (v4l) device. Hence, this implementation can take a real A/V signal as an input, if the video capture card in use has a driver for v4l.

The solution for mixing the A/V signals is to use customized JMF renderers for audio and video. All renderers know whether they are acting as a video mixer or if they are just passing data to the video mixer. Then, by using static object pointers, the renderer acting as a video mixer mixes the data coming from its input and the data obtained from the videoconferencing server's renderer.

Finally, the A/V signal is routed to the sound card and the video card. Then, the signal is converted into a signal, which the MPEG-2 encoder accepts, by using a hardware converter and fed into the encoder.

5 Results

The reference implementation was tested in Otadigi with real broadcasting equipment. The server was run in a Linux PC with AMD Duron 1.3 GHz processor and 1024 MB of memory. The size of the server byte code was small, only 327 kB. The server was connected to the MPEG-2 encoder through an Easy View signal converter, which converted the RGB signal from the computer to a SVHS signal.

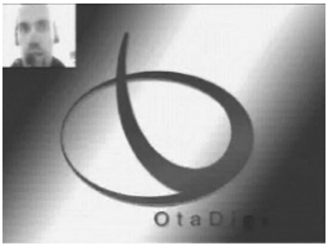


Fig. 4. Reference implementation screenshot

A videoconferencing connection was established between a simple SIP client with videoconferencing features and the server. The main problems were poor audio and video quality, and large delay between an active videoconferencing connection and the first video frame rendered, which was 9.8 s. In addition, the delay between the first video packet sent by the client and the first video frame rendered was 3.4 s. The delay was large because of JMF, which is not capable of handling delay-sensitive real-time connections. However, the feature was working, as shown in Figure 4, which was obtained from a commercial digital TV receiver. Moreover, the delay problem will be solved, if a commercial videoconferencing server is used.

6 Conclusion

Viewer participation in TV programs through videoconferencing is a new feature proposed in this paper. The feature can be implemented by adding two elements into TV broadcast system: videoconferencing server and a video mixer.

The feature has many advantages, which make it an interesting new feature for future TV broadcast systems. First, it is easy to implement, which is shown in this paper by developing a reference implementation. Second, it is a relatively cheap feature for both broadcast company and the viewers, as it uses common videoconferencing technology. Third, it enhances existing concepts using participation as well as enables the creation of new concepts.

The main problem with this feature is the lack of viewers with videoconferencing devices. Since this feature relies a lot in the accessibility of the feature, it is very important that most of the viewers are able to use it. The future looks good, although it is not sure how long does it take before the number of videoconferencing clients is high enough.

Interesting issues for further studies include integration with digital TV's interactive applications, and development of advanced features in set-top box providing videoconferencing client.

References

1. Wu, C.-H., Irwin, J. D.: Emerging multimedia computer communication technologies. Upper Saddle River, NJ: Prentice Hall (1998)
2. Eriksson, H.: MBONE: the multicast backbone. Communications of the ACM, Vol. 3, Issue 8 (1994) 54–60
3. Greenhalgh, C., Bowers, J., Walker, G., Wyver, J., Benford, J., Taylor, I.: Creating a live broadcast from a virtual environment. Proc. of the 26th annual conference on Computer graphics and interactive techniques, Los Angeles, CA (1999) 375–384
4. Otadigi: www.otadigi.tv
5. Herrero, C., Cesar, P., Vuorimaa, P.: Delivering MHP applications into a real DVB-T network, Otadigi. Proceedings of the 6th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services, TELSIKS2003, Nis, Serbia and Montenegro (2003) 231–234
6. Handley, M., Rosenberg, J., Schooler, E., Schulzrinne, H.: SIP: Session Initiation Protocol. IETF, RFC 2543 (1999)
7. International Telecommunication Union: Packet based multimedia communication systems. Recommendation H.323, Telecommunication Standardization Sector of ITU, Geneva, Switzerland (1998)
8. Hicks, J. A., Chappell, M. A.: Consumer interactive TV: what comes after the digital set-top box/TV combination? in Proc. of the 1st International workshop on Community Networking Integrated Multimedia Services, San Francisco, CA (1994) 71–74
9. Dalgic, I., Fang, H.: Comparison of H.323 and SIP for IP Telephony Signaling. in Proc. of Photonics East, Boston, Massachuttes (1999) 106–122
10. Karlsson, G.: Asynchronous transfer of video. IEEE Communications Magazine, volume 34, issue 8 (1996) 118–126
11. Rauhala, J., Cesar, P., Peisa, P., Vuorimaa, P.: Platform independent SIP client for consumer devices. in Proc. of the 7th IASTED International Conference on Internet and Multimedia Systems and Applications, Honolulu, Hawaii (2003) 552–557

Application Layer Multicast with Proactive Route Maintenance over Redundant Overlay Trees

Yohei Kunichika, Jiro Katto, and Sakae Okubo

Department of Computer Science, Waseda University
{yohei,katto}@katto.comm.waseda.ac.jp, sokubo@waseda.jp

Abstract. In this paper, an efficient algorithm to look for backup parents in preparation of parent leaving is proposed for application layer multicasting whose topology is constituted in the shape of a tree from a single source node. In most conventional methods, each child node starts searching for its new parent after its parent node leaves from a multicasting tree. This reactive operation often causes long interruption period. In our proposal, each node holds its parent candidate proactively over redundant overlay trees. Proactive route maintenance leads to smooth switching to a new parent after node leaving and failure, and redundant structure of a multicasting tree avoids exhausting search of a backup parent. Computer simulations are also carried out and effectiveness of the proposed approach is verified.

1 Introduction

Internet broadcasting has attracted attention of many researchers since the advent of IP multicasting [1]. The IP multicasting is an effective mechanism that can completely eliminate redundant data delivery to the multiple subscribers. However, the IP multicasting suffers from its quite slow deployment in the Internet due to inefficient support of native IP multicasting by the routers of current commercial ISPs. Application layer multicast (ALM) or overlay multicast emerges as an alternative to the IP multicasting. It enables packet multicasting delivery in an application layer without changing any network infrastructure of the current Internet. Instead of extending router functions, each end host receives a packet, replicates and forwards it to the next end hosts on an overlay network.

The most active research area about the ALM is a design of routing protocols [2]–[13]. There are several measures to evaluate effectiveness of the routing protocols as follows: (a) quality of the data delivery path, that is measured by stress, stretch and node degree parameters of the overlay multicast tree against native IP multicasting, (b) robustness of the overlay, that is measured by the recovery time to restore a packet delivery tree after abrupt end host failures, and (c) control overhead, that represents scalability of the protocols against large number of receivers.

One of the unavoidable problems of the ALM is that end hosts have to reconstruct the overlay network after a node leaves the multicast session or fails. In

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 306–313, 2004.

IP multicast, because the non-leaf nodes in the delivery tree are routers, we do not have to take into account the preceding problems. However, in the ALM, the non-leaf nodes are end hosts. End hosts are free to leave the multicast session, hence it is important to restore the packet delivery tree in these cases. Nevertheless, although quite a lot of routing protocols are proposed, most researchers focused on reactive restoration of a delivery tree. That is, end hosts start to search for its new parent after its old parent node departures.

Some researchers considered proactive approaches, in which backup routes are maintained before the parent departure happens. In Probabilistic Resilient Multicast (PRM) [12], each host chooses a constant number of other hosts at random and forwards data to each of them with a low probability. This operation indirectly contributes to backup route maintenance. However, PRM generates extra data overheads that cannot be negligible for some applications such as live media streaming. Another proactive approach called Yang's approach in this paper employed control packets instead of data packets [13]. In this approach, backup routes are calculated proactively whenever a node leaves or joins. When a node leaves, a backup route previously calculated is immediately applied and next backup routes are updated. When a new node joins, next backup routes are calculated without activating the previous backup route. Calculation of the backup routes is lead by two nodes; the parent node and the grandparent node of the leaving/joining node. A problem of this approach is that, when a node cannot form a backup route due to degree constraints of its children nodes, it employs grandchildren and below until a node without the degree constraint will be found. Another problem is that two nodes have to be involved in the backup route calculation whenever leave/join events happen. We therefore propose a novel proactive approach that avoids the degree limitation problem by forcing each node to prepare a redundant route for backup. Each host communicates with its grandparent and registers its backup parent information. When a node leaves the overlay network, it is guaranteed that all of its children can connect to their backup parents at worst in the same layer. In addition, the number of nodes to be engaged in the backup route update will be reduced to unity as shown later. The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 proposes the proactive backup route maintenance mechanism. Section 4 provides results of performance evaluation. Finally, Section 5 concludes this paper.

2 Related Work

Most of the application layer multicast protocols focus on how to construct an efficient multicast tree. Narada [2] and ALMI [3] are mesh-first protocols that were designed for small groups. Scattercast [4] is another mesh-first protocol for larger groups, which utilizes a set of proxies to which end hosts register. Yoid [5], Overcast [6] and Peercast [7] are tree-first protocols for larger groups. Bayeux [8] and CAN-based multicast [9] utilize structured P2P routing known as the distributed hash table (DHT) algorithm. ZIGZAG [10] and OMNI [11] are recently reported ALM protocols. ZIGZAG organizes a multi-layer hierarchy of bounded-size clusters of peers and constructs a multicast tree for each peer to receive contents from a "foreign" head. This procedure avoids occurrence of network bottleneck and keeps small end-to-end delay. OMNI presents a decentralized and adaptive update mechanism of an overlay network to minimize average-latency to the entire hosts with degree constraints. However, for node leaving and failures, these protocols employ reactive actions, i.e. they start to find new routes after the node departure happens. Therefore, long interruption period sometimes occurs in children nodes, which severely degrades streaming performance.

On the other hand, PRM [12] uses a proactive approach of reconstructing overlay network. Randomized forwarding seems to be effective in some situation, but this scheme may generate huge traffic of backup route maintenance. Yang's approach [13] forces each non-leaf host to pre-calculate a backup parent for each of its children. In this scheme, recovery time to restore overlay network is cut down with appropriate amount of control packets. However, in this approach, pre-calculation of a backup route sometimes consumes heavy computational cost because degrees of upper layer nodes are usually filled up. A node that seated in upper layers in the overlay cannot find its backup route until going down to lower layer nodes having no degree constraint.

3 Proposed Method

Construction of an overlay network is similar to constructing a degree-constrained spanning tree. Each end host is restricted to have children that receive data from the host, because the bandwidth is limited between the host and its children. It is known that the degree-constrained minimum spanning tree problem is an NP-complete problem, and many researchers had tried to alleviate the problem with some heuristics. However, instead of focusing on this problem, this paper pays attention to how to reconstruct a feasible spanning tree that can recover quickly in the case of node leaving and failures.

3.1 Reactive Reconstruction of an Overlay Network

When a host "leaves" an overlay network, it can send a message to inform affected nodes of its leaving the network. When a host suddenly "fails", it cannot send a message, so affected nodes have to detect this failure by some sort of heartbeat mechanism. We first take up Peercast [7] as an example of the reactive approaches. It proposed some recovery processes after a node leaves; root, root-all, grandfather, and grandfather-all. Among these, we choose the "grandfather" policy for comparison purpose because its performance is most graceful overall. In this grandfather policy, when a node leaves, the children of the leaving node contact their grandfather. If a node fails, the children of the node contact the source node rooted at the overlay tree because they cannot recognize their grandfather. The main task is to find a new parent for each affected child as quickly as possible. However, especially in the node failure phase, it takes longer time because each affected node searches for its new parent by contacting the rooted node in the tree, that might be quite far from the affected node. Furthermore, when the upper layer nodes are filled up in degrees, backup parent finding operation has to be repeated to reach the node that might be located at the lower layer (possibly, a leaf node).

3.2 Proactive Route Maintenance over Redundant Overlay Trees

We therefore apply a proactive approach in order to reduce the time of restoration of an overlay network. It is most important that we construct an overlay tree without each host maximizing its out-degree. Total out-degree may be calculated by the bandwidth of the connection of an end host divided by the media playback rate. If an end host has total out-degree = n, it can have n children in the previous work. However, in our scheme, we force an end host with total out-degree = n to have n-1 children only. This limitation simplifies backup route calculation, in which parent finding operation is completed at the children layer, and contributes to overhead reduction. Let us explain an example in detail. Firstly, new node participation process is carried out as follows. In Fig. 1, it is assumed that total out-degree of each node is equal to 4. In the previous work, when new node 8 requests to connect to node 1, node 1 accepts node 1 as its child because its degree is not filled. However, in our proposal, node 1 refuses the request because the rest of degree of node 1 is only one. As a result, node 8 is redirected to node 2. Next, backup route calculation is carried out by the parent node as follows. In Fig. 2, when node 8 connects to node 2, node 2 checks its children list. Since node 2 has three children, when node 2 leaves, node 1 cannot accommodate all the children of node 2 due to its degree constraint. Therefore, node 2 sends the children list to node 1. Node 1 then measures a round trip time to each grandchild, and informs the fastest node that node 1 will be its backup parent. Node 1 also informs the other nodes that faster nodes will be their backup parents. In Fig.2, node 5 is the fastest, node 6 is the second and node 8 is the last. Therefore, node 5 chooses its grandparent, node 1, node 6 chooses node 5, and node 8 chooses node 6, respectively, as their backup parents. As a special case, when the children list of node 2 includes node 8 only (i.e. no other children exist), node 2 immediately informs node 8 that node 1 will be a

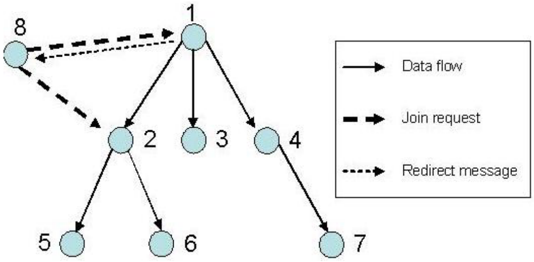


Fig. 1. An example of new node participation

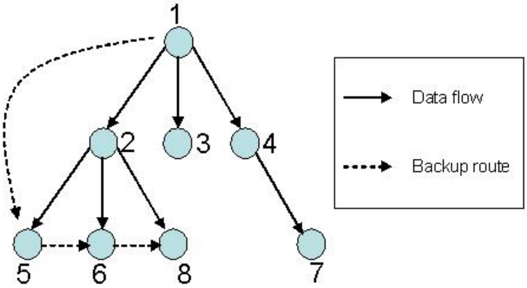


Fig. 2. An example of backup route selection

backup parent of node 8. This backup route calculation is also carried out whenever the node leaving/failure event happens similar to the node participation case. When a node leaves a network, the backup route which skips the vanished node is immediately applied and the new backup route calculation follows. Note that layers to which the backup route calculation is applied are limited at worst at the grandchild layer. It never goes down to the lower layers dissimilar to the previous approaches. Backup routes created above are certainly efficient as long as each host does not fully utilize its out-degree. However, it is possible that a host maximizes its out-degree by accommodating a new node after restoring an overlay tree. When this happens, a tree reconstruction procedure is invoked by the host itself in order to recover the route redundancy. Currently, this procedure is carried out by asking the children and below except the newly connected node whether their out-degrees are filled up. When an acceptable node is found, the newly connected node is moved to the acceptable node.

4 Performance Evaluation

We carry out computer simulations using ns-2 simulator [14]. We are mainly interested in the resilient performance that indicates how fast the overlay tree can be restored and in the reduction of control overheads thanks to provision of redundant routes. We compare our scheme with the promising reactive scheme, called the grandfather policy, described in Subsection 3.1. We also compare our scheme with Yang's approach, which is the previous proactive scheme proposed in [13]. Our simulation topology has 24 routers, in which four of them are domain-to-domain routers. End hosts randomly connect to one of the 20 routers except the four inter domain routers. Figure 3 represents a part of our topology. The number of hosts varies from 50 to 400. The link latencies vary from 10 ms to 100ms. The out-degree of each host is fixed at 4. The overlay tree is constructed at once after each experiment starts. Then end-hosts randomly join and leave the tree every 10 seconds.

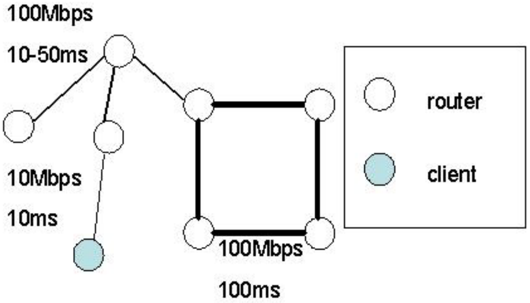


Fig. 3. Network topology used in computer simulations

4.1 Comparison of Recovery Latencies

Figure 4 compares recovery latencies of grandparent policy of the reactive approach, Yang's approach and the proposed backup route maintenance method. Recovery latency is the time after an affected node of a leaving node connects to a new parent and receives data packets from the parent. From this figure, we can recognize that the average recovery time of the reactive method is twice or more higher than that of the proactive methods when a node leaves, and 10 times higher when a node fails. This result is relevant because it is almost proportional to the average number of nodes contacted by children nodes of the leaving node. The proactive methods enable the affected nodes to connect to their backup parents immediately. This is common in both proactive methods, so their results are nearly equal. On the contrary, in the reactive method, the request may be rejected by the contacted node due to the degree constraint and be repeated until it will be accepted. Probability of this rejection becomes higher when each node contacts to an upper layer node, especially in the node failure case, where

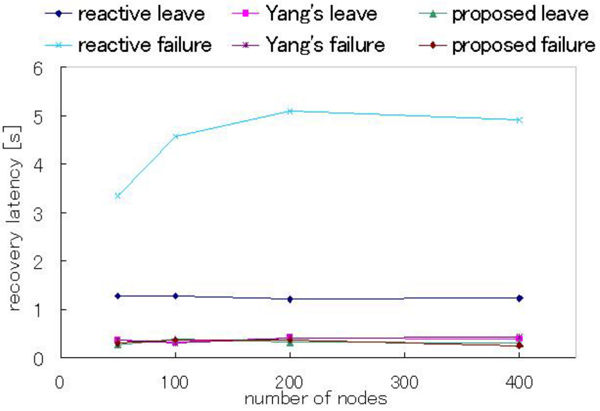
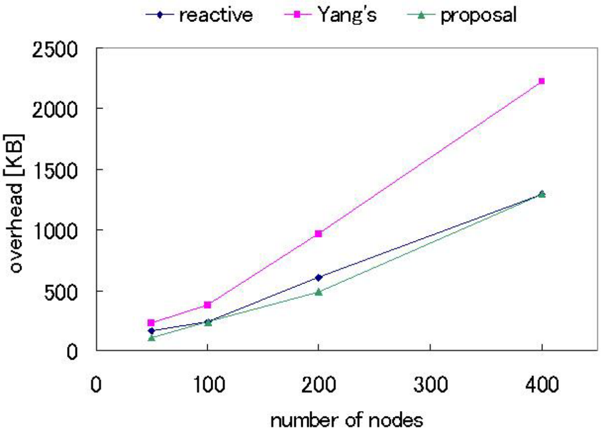


Fig. 4. Comparison of recovery latencies

affected nodes have to contact to the source node rooted at the overlay tree. As the number of end-hosts increases from 50 to 400, the recovery time of the reactive approach for node failure increases. This is because the height of an overlay tree becomes deeper on the average. On the other hand, recovery latencies of the proactive approaches are almost the same independently of the number of nodes because it is enough for each affected node to contact only one node.

4.2 Comparison of Control Overheads

Figure 5 compares total number of control packets of the reactive method, Yang's method and the proposed method. Control packets represent all signaling packets except data packets and heartbeat messages. From this figure, we can recognize that Yang's approach generates higher control packets than others. In addition to its possible iterative search problem, another reason is that it has to activate two nodes (parent and grandparent) for its backup route maintenance, because addition of a new node (child) sometimes invalidates previous backup routes that have to be updated by the grandparent. Note that our scheme activates a parent node only due to the route redundancy. We also observe that our scheme performs almost similar to the reactive scheme. This is because the reactive method generates more packets when iterative parent search is invoked. Thanks to the route redundancy again, our approach does not cause the iterative requests even if it generates excessive control packets for backup route maintenance. As a result, both overheads are almost the same.



5 Conclusions

This paper proposed proactive backup route maintenance over redundant overlay trees in order to enable smooth tree recovery and to reduce control overheads. Computer simulations were carried out, and it was verified that the recovery latencies were drastically reduced while the overhead of control packets was almost the same against the reactive approach. Furthermore, improvement to the existing proactive approach is also provided quantitatively. As future work, implementation of the proposed ALM system has to be evaluated along with the mathematical modeling.

References

1. Deering, S.: Host Extension for IP Multicasting, RFC 1112, Aug. (1989)
2. Chu, Y., Rao.G.S., Zhang, H.: A Case for End System Multicast, in Proceedings of ACM SIGMETRICS 2000, June. (2000)
3. Pendarakis, D., Shi, S., Verma, D., Waldvogel, M.: ALMI: An Application Level Multicast Infrastructure, 3rd USENIX Symposium on Internet Technologies and Systems, Mar. (2001)
4. Chawathe, Y., McCanne, S., Brewer, E.: Scattercast: An Architecture for Internet Broadcast Distribution as an Infrastructure Service, PhD Thesis, University of California, Berkeley, (2000)
5. Francis., P.: Yoid: Extending the Internet Multicast Architecture, http://www.icir.org/yoid/
6. Jannotti, J., Gifford, D., Johonson, K., Kaashoek, M., O'Toole, J.: Overcast: Reliable Multicasting with an Overlay Network, 4th Symposium on Operating Systems Design & Implementation, Oct. (2000)
7. Deshpande, H., Bawa, M. Garcia-Molina, H.: Streaming Live Media over Peers, Technical Report 2002-21, Stanford University, Mar. (2002)
8. Zhuang, S., Zhao, B., Joseph, A., Katz, R., Shenker, S.: Bayeux: An Architecture for Scalable and Fault-Tolerant Wide-Area Data Dissemination, ACM NOSSDAV 2001, June. (2001)
9. Ratnasamy, S., Handley, M., Karp, R., and Shenker, S.: Application-level Multicast using Content-Addressable Networks In Proceedings of NGC (2001)
10. Tran, D., Hua, K., Do, T.: ZIGZAG: An Efficient Peer-to-Peer Scheme for Media Streaming, in proceedings of IEEE INFOCOM 2003, Apr. (2003)
11. Banerjee, S., Kommareddy, C., Kar, K., Bhattacharjee, B., Khuller, S.: Construction of an Efficient Overlay Multicast Infrastructure for Real-time Applications, in proceedings of IEEE INFOCOM 2003, Apr. (2003)
12. Banerjee, S., Lee, S., Bhattacharjee, B. Srinivasan, A.: Resilient multicast using overlays, in proceedings of ACM SIGMETRICS 2003, June. (2003)
13. Yang, M. Fei, Z.: A Proactive Approach to Reconstructing Overlay Multicast Trees, in proceedings of INFOCOM 2004, March. (2004)
14. The Network Simulator -ns-2, http://www.isi.edu/nsnam/ns

Real-Time Rate Control Via Variable Frame Rate and Quantization Parameters

Chi-Wah Wong[1], Oscar C. Au[1], Raymond Chi-Wing Wong[2], and
Hong-Kwai Lam[1]

[1] Hong Kong Univ. of Science and Technology, Hong Kong
{dickywcw,eeau,eekwai}@ust.hk
[2] The Chinese University of Hong Kong, Hong Kong
cwwong@cse.cuhk.edu.hk

Abstract. Most of existing rate control schemes in the literature focus on the bit allocation under the assumption of constant frame rate. As a result, the distortion of consecutive frames varies in a great extent, which may result in low temporal quality (e.g. flickering effect). In this work, we design rate control with two control parameters (i.e. quantization factor and frame rate) to achieve better tradeoff between spatial quality and temporal quality in low complexity. Our optimization gives two suggested quantization step sizes. One is used to change the frame rate in order to allocate how many bits used in the current frame. Another one is used to change the quantization factor in order to quantize MBs to achieve target bit rates. The experimental results suggest that our scheme can achieve more consistent quality while keeping high spatial quality.

1 Introduction

Standard video systems, such as H.261/2/3/4 and MPEG, exploit the spatial, temporal and statistical redundancies in the source video. Since the level of redundancy changes from frame to frame, the number of bits per frame is variable, even if the same quantization parameters are used for all frames. Therefore, a buffer is required to smooth out the variable video output rate and provide a constant video output rate. The rate control is used to prevent the buffer from overflowing (resulting in frame skipping) or/and under-flowing (resulting in low channel utilization) in order to achieve good video quality. For real-time video communications such as video conferencing, it is more challenging as the rate control is required to satisfy the low-delay constraints, especially in low bit rate channels.

Many traditional rate control schemes (e.g. [1], [4], [5], [6]) adjust quantization parameters of the macro-blocks (MB) only. However, there are other control parameters to do the rate control (e.g. spatial and temporal resolutions). Some current research proposed the rate control schemes by adjusting another control parameter (e.g. frame rate)[2], [3]. By doing this, consistent quality can be obtained over frames. In other words, the flickering effect, caused by the fluctuation of spatial image quality between consecutive frames, can be reduced. Although the fact that the change of PSNR does not correspond to flickering completely,

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 314–322, 2004.

it is observed that the flickering effect can be reduced by keeping the image quality of each frame almost constant [3]. However, their schemes have large computational complexity and time delay although sub-GOP is used instead of GOP. In this work, we design a new rate control scheme with low complexity to have better tradeoff between spatial quality and temporal quality by changing both quantization parameters and frame rate.

In this work, we present a frame-rate controlled rate control scheme for encoders in real-time video communications. This work focuses on doing rate control for inter-coded frames (i.e. P-frame), which is used mostly in low-delay video communication. We first describe what rate and distortion models are being used. Based on these models, we minimize the distortion subject to the target bit constraint and minimize the bit rate subject to the target distortion in two optimizations. By Lagrange optimization, we obtain two formulas that indicate how to choose the quantization parameters and the frame rate to encode frames with high spatial quality and consistent quality.

This paper is organized as follows. In the following section, we describe the rate and distortion models. In section 3, we describe two optimizations with the target bit rate and the target distortion constraint. In section 4, our proposed rate control scheme is described. In section 5, the experiments are conducted to evaluate the performance. Finally, the conclusion is made.

2 Rate and Distortion Modeling

In DCT-based motion compensated video encoders, the current video frame to be encoded is decomposed into 16x16 macroblocks (MB). Motion estimation and compensation are applied to give the residue MB, each of which is divided into sixteen 4x4 blocks and discrete cosine transform is applied to the 4x4 residue blocks in H.264 [8]. After that, the DCT coefficients within a block are quantized, zigzag scanned and encoded with variable length coding in general. The number of encoded bits and distortion of a given MB are observed to be dependent on the MB's quantization step size Q and the standard deviation σ of the residue MB.

By the property of the motion prediction process, the pixel values of the residue MB tend to have a characteristic distribution of Laplacian probability density function with standard deviation σ. Based on this, we use the following quadratic rate model R_i and MSE distortion model D_i [1] of the i-th MB in our optimization.

$$R_i = A(K\sigma_i^2/Q_i^2 + C) \tag{1}$$

$$D_i = a\alpha_i^2 Q_i^2 \tag{2}$$

where A is the number of pixels in MB, σ_i is the standard deviation of the i-th MB, Q_i is the quantization step size of the i-th MB, K is the rate model parameter, C is the overhead rate, a is the distortion model parameter, α_i is the weight of the i-th MB.

This rate model is valid at low bit rates whereas the distortion model is valid for uniform quantizers of uniform data distribution with $a = 1/12$ and $\alpha_i = 1$. Although these two models are not very accurate, it is quite suitable for real-time rate control because these models are simple and efficient to reduce the complexity of updating model parameters. These models are used from the optimizations in the following section.

3 Optimization

In this section, we derive formulas for two optimal quantization step sizes Q^*. One is to minimize MSE distortion subject to the target bit constraint. Another is to minimize bit rate subject to the target distortion constraint. The same rate model and distortion model just described are used for these two optimizations.

3.1 Constant Target Bit Rate, *B*

The quantization step sizes are chosen based on the following optimization formula. The original problem is

$$Q_1^*, Q_2^*, ..., Q_N^* = \underset{\substack{Q_1,...Q_N \\ \sum_{i=1}^{N} R_i = B}}{\arg\min} \frac{1}{N}\textstyle\sum_{i=1}^{N} D_i \tag{3}$$

where Q_i^* is the optimal quantization step size of the i-th MB, $R_i = A(C + K\sigma_i^2/Q_i^2)$ is the estimated number of bits for i-th MB, , B is the target number of bits for the frame, N is the number of MB of the frame.

By using Lagrange optimization, the problem becomes

$$Q_1^*, Q_2^*, ..., Q_N^*, \lambda^* = \underset{Q_1,...Q_N,\lambda}{\arg\min} \{\tfrac{1}{N}\sum_{i=1}^{N}(a\alpha_i^2 Q_i^2) + \lambda[\sum_{i=1}^{N} A(K\tfrac{\sigma_i^2}{Q_i^2} + C) - B]\} \tag{4}$$

The expression (found in [1]) for the optimization quantization step size Q_i^* in the bit-constraint optimization is

$$Q_1 : Q_i^* = \sqrt{\frac{AK}{B - ANC}\frac{\sigma_i}{\alpha_i}\textstyle\sum_{k=1}^{N}\alpha_k\sigma_k} \quad i = 1, 2, ..., N. \tag{5}$$

This quantization step size is used to quantize the coefficients in our rate control to achieve the desired target bit rate B.

3.2 Constant Target Distortion, *D*$_0$

Similar to the approach to the bit-constraint optimization, the constant-distortion problem is

$$Q_1^*, Q_2^*, ..., Q_N^*, \lambda^* = \underset{Q_1,...Q_N,\lambda}{\arg\min} \{\sum_{i=1}^{N} A(K\tfrac{\sigma_i^2}{Q_i^2} + C) + \lambda[\tfrac{1}{N}\sum_{i=1}^{N}(a\alpha_i^2 Q_i^2) - D_0]\} \tag{6}$$

where D_0 is the target distortion of the current frame.

The expression for the optimization quantization step size Q_i^* in the distortion-constraint optimization is

$$Q_2 : Q_i^* = \sqrt{\frac{\sigma_i N D_0}{a\alpha_i \sum_{k=1}^{N} \alpha_k \sigma_k}} \quad i = 1, 2, ..., N. \tag{7}$$

This quantization step size Q_2 is used as a reference to change the frame rate such that the desired target bit rate can be obtained. In the Algorithm section, the details of how to use this step size are described.

Eq. (5) and Eq. (7) have common behavior.

1. They (i.e. Q_i^*) both increases with σ_i. Based on the equations, they are directly proportional to $\sqrt{\sigma_i}$.
2. They depend only on one model parameter. Eq. (5) and Eq. (7) is dependent on K and a only respectively.
3. They both contain the term $S = \sum_{k=1}^{N} \alpha_k \sigma_k$. Once S is calculated in one equation, S can be re-used in another equation. This reduces computational complexity.
4. Eq. (5) is directly proportional to eq.(7) (i.e. $Q_1 = LQ_2$ where L is constant). They all are proportional to $\sqrt{\sigma_i/\alpha_i}$. $Q_1 = L_1\sqrt{\sigma_i/\alpha_i}$ and $Q_2 = L_2\sqrt{\sigma_i/\alpha_i}$ where L_1 and L_2 are constant. Once one quantization equation is used, another quantization equation can be obtained through the relationship equation (i.e. $Q_1 = LQ_2$) instead of the original equation. This also reduces the computational complexity, which is much smaller than the scheme in [2] and [3]. In [2] and [3], their schemes require to calculate the estimated histogram of the difference (HOD) between two consecutive frames each time and find quantization factors by multi-pass iterations. The computational complexity and time delay are still large although sub-GOP is used instead of GOP.

4 Rate Control Algorithm

In this section, we will propose our rate control algorithm. In our experiments, the first frame is intra-coded (I-frame) with a fixed quantization parameter. The following frames are of type P. This means that they are predicted from the corresponding previous decoded frames using motion compensation and the residue is obtained. First, we do the rate control in the frame layer. After that, we do the rate control in macro-block level, which is similar to TMN8.

4.1 Frame-Layer Rate Control

The encoder buffer size W is updated before the current frame is encoded.

$$W = max(W_{prev} + B' - Rm/F_0, 0) \tag{8}$$

where W_{prev} is the previous number of bits in the buffer, B' is the actual number of bits used for the encoded previous frame, R is the channel bit rate (bit per

Table 1. Actual frame, skipped frames and relationship between j and m. (Original frame rate: 30fps)

j	1	2	3	4	5	6
m	1	2	3	4	6	12
Skipped frames	0	1	2	3	5	11
Frame rate (fps)	30	15	10	7.5	5	2.5

sec), F is the initial frame rate (frame per sec), and m is the control parameter to change a frame rate.

After updating the buffer size, if W is larger than or equal to the predefined threshold $M(= R/F)$, the encoder skips encoding the frames until W is smaller than M. This means that buffer overflow will not occur at the cost of frame skipping. Same as [2] and [3], the frame rate is only updated for a period of 12 frames. The possible frame rates (i.e. 30, 15, 10, 7.5, 5, and 2.5) of each period are obtained from the look-up Table 1, which is the same in [2] and [3]. When frame rate is changed, the current frame rate will be stepped down or up by 1. For example, if the previous frame rate is 10 fps, possible next frame rate is 7.5, 10 or 15 fps. Human is sensitive to large sudden change in frame rate (e.g. 2 or more steps). Within the period, the frame rate remains unchanged. For the first frame of each period,

1. Compute all of σ_i
2. Compute $S_1 = \sum_{k=1}^{N} \alpha_k \sigma_k$
3. Compute the optimal Q, $Q_i^* = \sqrt{\frac{\sigma_i N D_0}{a \alpha_i S_1}}$ where D_0 is the average distortion of frames from previous period.
4. Estimate the number of bits required per frame for small change in distortion $\hat{R} = \sum_{i=1}^{N} A(K \frac{\sigma_i^2}{Q_i^{*2}} + C)$
5. Change the frame rate (F_0: initial frame rate)
   Define $\hat{R}_u = (m+1)R/F_0$ and $\hat{R}_l = (m-1)R/F_0$
   – If $\hat{R} > \hat{R}_u$, then $j = j + 1$ and find the corresponding m from look-up Table 1 (decrease the frame rate)
   – If $\hat{R} < \hat{R}_l$, then $j = j - 1$ and find the corresponding m from look-up Table 1 (increase the frame rate)
6. Calculate the target bit of each frame $B = Rm/F_0 - \varDelta$
   where

$$\varDelta = \begin{cases} W/F & W > 0.1M \\ W - 0.1M & , otherwise \end{cases} \tag{9}$$

Step (1) calculates σ_i of all the MBs in a frame. Step (2) computes S which is used directly by two models (described in behavior (3)). The purpose of step (3) is to estimate Q_i^* such that the distortion of the current frame can be similar to the average distortion of frames from its previous period. In Step (4), the number of required bits per frame $\hat{R}$ with Q_i^* is estimated. $\hat{R}$ indicates how many bits to be used in the current frame to achieve similar distortion from its previous period. If $\hat{R}$ is very large, this means that each frame of the current period should need more bits. In order to do this, frame rate should be decreased. When $\hat{R} > \hat{R_u}$, the current frame does not have enough bits to be encoded to obtain similar distortion by using the frame rate just used previously. Therefore, frame rate should be stepped down for the purpose of small variation of distortion. On the other hand, when $\hat{R} < \hat{R_l}$, current frame needs less bits and frame rate should be stepped up for the same purpose. The frame rate keeps the same when $\hat{R_l} < \hat{R} < \hat{R_u}$.

We know that the frame rate cannot be changed in a greater extent because human is sensitive to the rapid change in frame rate. This means that the possible changes of current frame rate can be chosen as unchanged, stepped down by 1 or stepped up by 1 (i.e. j, $j+1$, or $j-1$).

The rationale of setting of $\hat{R_u}$ and $\hat{R_l}$ is described as follows. If the frame rate is not needed to be changed, the current target bit per frame should be similar to mR/F_0 with $\Delta = 0$. As the possible frame rate is $(m-1)/F_0$, m/F_0 and $(m+1)/F_0$, the possible target bits per frame for the current frame rate is $(m-1)R/F_0(=\hat{R_l})$, mR/F_0 and $(m+1)R/F_0(=\hat{R_u})$ respectively. If $\hat{R}$ of the current frame is within the range between $\hat{R_u}$ and $\hat{R_l}$, the frame rate should not be updated in order to avoid frequent updates of frame rate. Besides, we should not make any assignment on the bound $\hat{R_u}$ and $\hat{R_l}$ in a great difference with $\hat{R}$ (e.g. $\hat{R_u} = (m+2)R/F_0$ and $\hat{R_l} = (m-2)R/F_0$). If we define the bound $\hat{R_u}$ and $\hat{R_l}$ with a large difference with $\hat{R}$, the frame rate may not be changed rapidly and distortion varies in a great extent to, which human is sensitive. On the other hand, if we define the bound $\hat{R_u}$ and $\hat{R_l}$ with an extremely small difference with $\hat{R}$, it is unnecessary that the frame rate is updated frequently for different periods.

4.2 Macro-Block Layer Rate Control

In H.264, QP is required to be given to do the mode selection. For simplicity, the coding mode of each MB is determined based on the average QP of the previous frame. By doing this, σ_i of all MBs in the current frame can be obtained for our MB Layer rate control in a similar way of TMN8 [1].

Algorithm 1 Macro-block Layer Rate Control

```
for each P-Frame do
    $B_1 = B$ and $N_1 = N$;
    for each MB ($i = 1$ to $N$) do
        Compute $Q_i^* = \sqrt{\frac{AK}{B_i - AN_iC}\frac{\sigma_i}{\alpha_i}S_i}$ based on Eq. (5)
        Use the corresponding QP to quantize and encode the $i$-th MB
        Compute $S_{i+1} = S_i - \alpha_i\sigma_i$, $N_{i+1} = N_i - 1$ and $B_{i+1} = B_i -$
        ($actual\ current\ bits$)
        Update the rate model parameters $K$ and $a$ based on Eq. (1) and Eq. (2)
        Update the overhead parameter $C$ accordingly
    end for
end for
```

5 Experimental Results

We implemented the rate control scheme in a JVT JM 4.1 version [7]. In the following experiments, we compare the proposed rate control algorithm with TMN8 [1] and a rate control [3], called Song. The Song scheme also controls the frame rate but uses different implementation from our scheme. The first frame was intra-coded (I-frame) with QP=31, several frames were skipped after the first frame to decrease the number of bits in the buffer below target buffer level $M = R/F$ and the remaining frames were all inter-coded (P frames). This means that the number of skipped frames is the same in TMN8, Song and our proposed schemes (for fair comparison). Afterwards, they use their own scheme. The proposed algorithms, TMN8 and Song were simulated on some QCIF test sequences with initial frame rate of 10Hz and various target bit rates. Here are the test conditions. The MV resolution is at 1/4 pel. Hadamard is "OFF". RD optimization is "OFF". Search range is "± 16". Restrict search range is "0". Reference frames is "1" and symbol mode is "UVLC".

Table 2 shows the actual encoded bit rates achieved by TMN8, Song and the proposed rate control. They verify that these rate control methods can achieve the target bit rates. The error between target bit rate and actual bit rate is below 0.2%. For fair comparison, similar frame rates (7.5 - 15) are shown for test sequences observed from Table 3. In TMN8 scheme, frame rate is constant (i.e. 10 fps). In Song scheme, parameters (e.g. w_h and T) are set to achieve similar frame rate obtained by our proposed scheme. Table 4 shows the comparison of PSNR of the reconstructed pictures for TMN8, Song and the proposed rate control. The proposed rate control has similar to PSNR obtained in TMN8 and Song. To have better tradeoff between spatial and temporal quality, the average frame rate increases with bit rate in general because when the bit rate increases, there exists enough number of bits to encode frames themselves. Due to the sufficient of bits, there is a high tendency to increase the frame rate and have better temporal quality. In addition, it is observed that PSNR decreases with frame rate at the same target bit rates in general. This tendency can be obtained in Table 3 and Table 4 except "Sil48". In "Sil48", the correlation between consecutive frames is high as the frame rate is high and the frame time is shortened that two consecutive frames are quite close together. As a result, good prediction can be made that the residue has less bits and the quality may be better.

Table 4 also shows the PSNR variation over frames in TMN8, Song and the proposed rate control. PSNR variation (i.e. more consistent quality over frames) in the proposed rate control over TMN8 and Song is the smallest. And, Song scheme has smaller variation than TMN8 scheme. This is due to variable frame rate of the proposed rate control. When large distortion is desired, the frame rate is adjusted to be larger. Otherwise, the frame rate is adjusted to be smaller. As a result, the PSNR variation will be smaller to maintain consistent quality over frames. Although the variance of PSNR is not an exact measure of the flickering effect, it is fair to say that the flickering effect can be reduced by smaller variance of PSNR in slow-motion video (e.g. Akiyo, M & D and Silent) [3]. Then our scheme can also reduce the flickering effect. Fig 1 shows comparison of PSNR

Table 2. Comparison of bit rate achieved by TMN8, Song and the proposed rate control

Test Name	Video Sequence	Target Bit (kbps)	Encoded bits		
			TMN8	Song	Proposed
Aki24	"Akiyo"	24	24.03	24.04	24.02
Fmn72	"Foreman"	72	72.07	72.04	72.05
Ctg72	"Coastguard"	72	72.07	72.09	72.08
Mad24	"M & D"	24	24.03	24.03	24.02
Sil48	"Silent"	48	48.04	48.03	48.03
Stf256	"Stefan"	256	256.24	256.21	256.17

Table 3. Comparison of average frame rate for TMN8, Song and the proposed rate control

Test Name	Frame Rate (fps)		
	TMN8	Song	Proposed
Aki24	10	9.70	9.90
Fmn72	10	11.50	11.20
Ctg72	10	9.50	9.70
Mad24	10	8.10	7.81
Sil48	10	10.20	10.30
Stf256	10	13.50	13.50

Table 4. Comparison of average PSNR for TMN8, Song and the proposed rate control

Test Name	PSNR(dB)			Var in PSNR (dB)		
	TMN8	Song	Proposed	TMN8	Song	Proposed
Aki24	38.84	39.00	39.22	1.4091	1.1384	0.8192
Fmn72	34.12	33.95	33.93	2.8989	2.5073	2.1073
Ctg72	31.10	31.15	31.20	0.5277	0.4811	0.2700
Mad24	35.85	36.65	36.70	2.5103	1.5774	1.2602
Sil48	34.54	34.70	34.82	0.7082	0.6245	0.3986
Stf256	33.52	33.45	33.40	7.8176	2.3355	1.5185

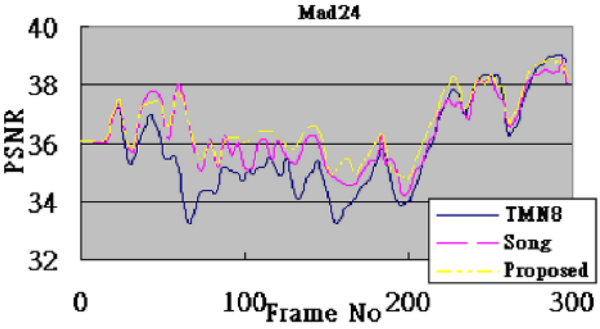


Fig. 1. Comparison of PSNR against frame number in "Mad24"

against frame number in “Mad24”. It is observed that the PSNR variation is small over frames in our proposed scheme. In some frames (e.g. around frame 100), the short-term PSNR variation is also small in our proposed scheme.

6 Conclusion

In this paper, we design a rate control with two control parameters (i.e. quantization factor and frame rate) to have better tradeoff between spatial quality and temporal quality in low complexity. Our optimization introduces two suggested quantization step sizes. One is used to change frame rate in order to allocate how many bits used for the current frame. Another one is used to change quantization factor in order to quantize MBs to achieve the target bit rates. The experimental results suggest that our scheme can achieve more consistent quality while keeping high spatial quality.

References

[1] J. Ribas-Corbera and S. Lei, “Rate Control in DCT Video Coding for Low-Delay Communications”, IEEE Trans. Circuits Syst. Video Technol., vol. 9, pp. 172–185, 1999

[2] H. Song, J. Kim and C.-C. J. Kuo, “Improved H.263+ Rate Control via variable frame rate adjustment and hybrid I-frame rate”, in Proc. IEEE Int. Conf. Image Processing(ICIP) 1998, vol. 2, pp. 375–378, 1998

[3] H. Song and C.-C. J. Kuo, “Rate Control for Low-Bit-Rate Video via Variable-Encoding Frame Rates”, IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 512–521, 2001

[4] H. J. Lee and T. H. Chiang and Y. Q. Zhang, “Scalable Rate Control for MPEG-4 Video”, IEEE Trans. Circuit Syst. Video Technol., vol. 10, pp. 878–894, 2000

[5] Z. He and S. K. Mitra, “A Linear Source Model and a Unified Rate Control Algorithm for DCT Video Coding”, IEEE Trans. Circuit Syst. Video Technol., vol. 12, pp. 970–982, 2002

[6] C.-W. Wong, O. C. Au, B. Meng and H.-K. Lam, “Perceptual Rate Control for Low-Delay Video Communications”, in Proc. IEEE Int. Conf. on Multimedia and Exco (ICME), vol. 3, pp. 361–364, 2003

[7] “JVT JM4.1”, ftp://ftp.imtc-files.org/, 2002.

[8] “H.264/MPEG-4 Part 10 Tutorials”, http://www.vcodex.fsnet.co.uk/h264.html

The Structure of Logically Hierarchical Cluster for the Distributed Multimedia on Demand

Xuhui Xiong, Shengsheng Yu, and Jingli Zhou

Department of Computer Science and Engineering, Huazhong Univ. of Sci. & Tech.
xhxiong@wtwh.com.cn, {ssyu,jlzhou}@mail.hust.edu.cn

Abstract. In this paper, we propose the structure of the logical hierarchical cluster for the distributed multimedia on demand according to block the nodes cache. The structure of the logical hierarchy is lazily maintained by all members in the decentralized manner. The global load balance is achieved as the root of each LHC is mapped into the system randomly. The LHC is expanded dynamically with the two application policies. The local load balance is applied to fine-tune the load of nodes of different LHCs within a local region. The integration of topology with resource management simplifies the server selection largely.

1 Introduction

In this paper, we focus on developing the logical hierarchical cluster (LHC), a logical decentralized architecture, for the distributed multimedia on demand (DMoD) application. Because the aggregated bandwidth requirement of the geographically distributed clients is increasing beyond the capacity of the network backbone, the DMoD scheme is proposed to distribute the load of the interactive MoD service. The network topology and the resource management are the two main components involved in the design of the DMoD architecture. The nodes with the limited storage capacity and streaming bandwidth are often arranged into a physically hierarchical topology [1] or a logically hierarchical management domain [2] which impact the resource management critically.

The DMoD system must possess the optimality of resource utilization and scalability. The methods of load balance mostly are the block placement, replication and server selection to maximize the resource utilization. Due to the limited resource of individual node and the huge size of the continuous multimedia (CM) objects, the block placement partitions a object into the blocks which are mapped on different nodes. Replication can contribute to the load balance and alleviate the stress of network backbone for the scalability [3]. Server selection is used to find an appropriate node from several caches for the client request.

In order to provide the scalability for the DMoD application, the nodes are organized into the physical network hierarchy [1] according to the underlying telecommunication infrastructure or the logical management hierarchy [2] in terms of geographical/organization regions. Nevertheless, for reasons of the variable user access behavior in the distributed environment, it is still difficult to

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 323–330, 2004.

accurately acquire the dynamic information about the load state of nodes as well the object location for the static structure.

Derived from the above two hierarchies, we construct a dynamic LHC for each block which is expanded along with the client request in the run time. In fact, all nodes are widely deployed at the network edge with the high-speed link. By the distributed hash table (DHT) [4], the physical path of node pairs is replaced by the logical link and each node maintains a small size of application routing table for the message transmission. Different to the above two hierarchies, the node joins into a LHC in the light of the block which it caches.

All members of a LHC are lazily involved in the maintenance of that LHC in the decentralized manner. Because the root of each LHC is randomly mapped into the system, we can achieve the global load balance. The logical structure will be expanded by the two difference replication policies. In particular, the early replication method is used to cache the block frequently requested in short term. As the current LHC cannot serve a new request, the instantaneous replication will be triggered. In addition, the local load diversion will be used for the load balance of nodes belong to different LHC in a local region. Guaranteed by the above load balance methods, the client will select the closest node with idle capacity by flooding the request in LHC.

This paper is organized as follow. In section 2, we describe the design of DMoD. In section 3, we evaluate the resource utilization of system. In section 4, we conclude the paper and discuss the future work.

2 Mechanism

In this section, we construct the LHC for each block to replace the physical network hierarchy and the logical management hierarchy. The structure of a LHC is dynamically expanded with the client access behavior and the node load state in the decentralized manner to simplify the server selection.

2.1 Structure of LHC

As the storage and delivery granularity, the CM object is partitioned into blocks. Based on the DHT location mechanism, a block is initially mapped on the node with a nodeID closest to the blockID by the DHT [4]. Each block will be cached to reduce the streaming traffic on the backbone caused by user accessing the remote server or enable the load balance of different nodes.

All the nodes, which cache the identical block, are formed into a scalable LHC which is dynamically expanded in the decentralized manner (see the section 2.3). On the other hand, a node might belong to numbers of LHCs at the same time. That is, the connectivity of nodes within a LHC is based on the block instead of the geographical/organized region or the network topology.

In the logical structure, the lowest layer comprises all members, namely the whole caches. Members in each layer of the LHC are partitioned into clusters. A member exists in only one cluster at any layer. The head of each cluster is the member with a nodeID numerically closer to the blockID than others in that cluster. The head of a cluster at layer i joins into the higher layer $i+1$. Specially, L_0 layer includes k members at least for data availability.

Each member maintains three tables for its siblings, parents and children respectively. Namely, the maintenance of a LHC is in the soft-state manners. The entries of tables are arranged in ascending order of IP delay. With the DHT mechanism, the height of a LHC is $O(\log_b N)$. Since the nodes and blocks are uniformly mapped onto a huge ID space, the maximal average size of a single cluster is b members.

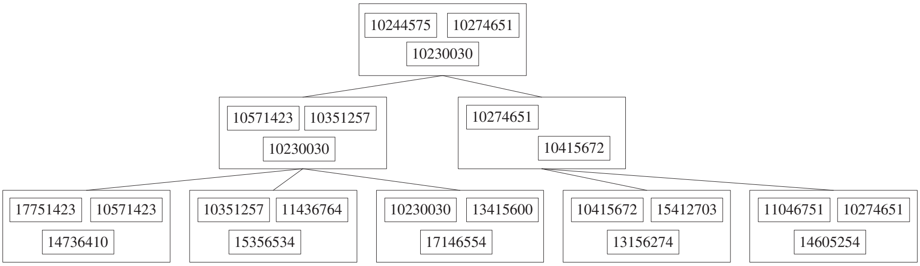


Fig. 1. A temporary 3-level LHC for block 10245326. Where, $k = 3$

2.2 Maintenance of LHC

A block's LHC can be expanded dynamically by the instantaneous replication and the early replication strategy (see section 2.3). In both cases, the new cache joining mechanism is very simple: the first cache encountered on the enquiry path is primary parent of the new cache. The new cache acquires the address of uncles and siblings from the primary parent and registers into these nodes.

A new cache might passively join into the lower clusters as head, in the process of the LHC refinement. The node rearranges its parents and siblings when the latency between itself and its parent and siblings changes significantly. Therefore, a node A emigrates to another cluster headed with node B, as following conditions are satisfied: (1)$\min(delay(A, B)) < delay(A, A_l)$; (2)$share(B, ID) < share(A, ID)$. A_l is the head of the original cluster, where $B \in A_{parent} \cup A_{sibling}$. ID is the blockID of this LHC and $share(x, y)$ is the length of prefix shared among x and y.

If a member actively leaves its clusters or it is dead due to failure, it will be removed by its parents and siblings directly. In addition, its children need rejoin one cluster headed with its sibling similarly to the above LHC adjustment process. If such cluster cannot be found (e.g., a node has only a single parent node), the children will route a message with blockID of this LHC to rejoin grandfather's cluster. Note that, this system is read-only and blocks can always be located via DHT-routing, so the nodes of LHC lazily exchange heartbeat with each other. More-over, the state information is piggybacked in the flooded request message to alleviate the control overhead.

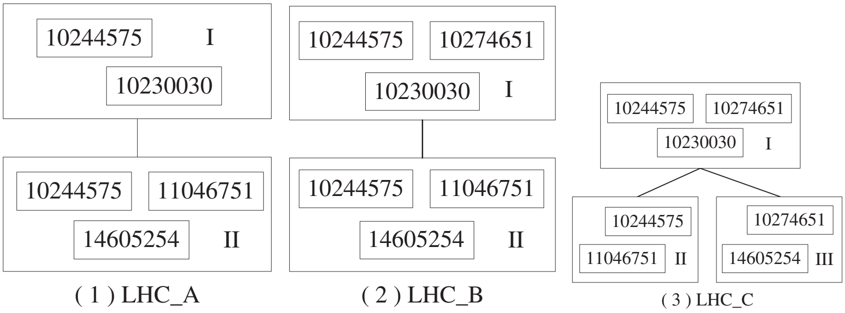


Fig. 2. Node 10274651 will joins the LHC_A of block 10245326. Firstly, it joins the highest cluster I headed with node 10244575 and results LHC_B. Next, node 14605254 in lower cluster II perceives node 10274651 is nearer than node 10244575 in the refinement process, it will move to the cluster III headed with node 10274651 and then results LHC_C.

2.3 Dynamical Expansion of LHC

In the CM object delivery process, we use two different replication policies to achieve the load balance further and alleviate the backbone stress.

Firstly, as the load of system is low, we use the early replication policy to actively replicate the block requested frequently in a short term. Based on the load-states of nodes (see section 2.5), each node of a LHC maintains two tables which log the admitted request for the first block during $T_{Current-(N-1)}$ to $T_{Current}$ and the successive blocks respectively from $T_{Current}$ to $T_{Current+(2N-2)}$. Whenever the node A finds there exist two or more requests for the same block from the identical proxy B, if both node A and B have more available resource than the threshold T_{pri}, then the block is replicated from A to B. Due to the access skewness of the CM objects, the statistic of cumulative client requests during the short term only triggers the replication of the blocks of popular objects. Moreover, the early replication policy can also avoid the contention for the network bandwidth between the replication overhead and normal delivery load during the peak time.

On the other hand, as the system load becomes heavy, if the nodes of LHC have no available resource, the instantaneous replication will be triggered. The request for a block forwarded by the proxy of client will arrive a leaf of LHC. The sub-set of targets for replication is composed of the intermediate nodes with available resource from that proxy to the leaf. In order to server the clients more widely, the instantaneous replication policy will place a cache from the farthest node of the sub-set which meets the IP latency constraint. The client's request is rejected by the LHC and the respective sub-set when the cache cannot be placed. If the servers are densely deployed on a local region (exceeds the size of neighbor set in the DHT), the client can retry to select an-other proxy for the sub-set of replication targets.

2.4 Local Load Diversion

Typically, the nodes deployed in the local region are increasing with the clients and belong to different LHCs. We use the local load diversion to finely tune the network load to improvement of resource utilization .A node and its neighbor set forms a virtual server group instead of statically partitioning a server group. The node exchanges the load-state information of with its neighbors periodically.

According to its load-state, each node will divert/receive load to/from neighbors. Note that, if the number of nodes deployed on a local region is large than the size of neighbor set, the two virtual server groups might partly intersect. Thus, the load diversion of that local region is transferable. Load diversion includes diversion trigger, source block selection and target node selection policy.

- Each node has a diversion threshold T_d and a diversion acceptance threshold T_a. If a new request makes load $L_s(t)$ maintained by the node s exceed T_d $(1 \leq t \leq 2N - 1)$, then the load diversion is triggered.
- The total load held by a node at each unit time is $\sum L_S^{S_i}(t)$, S_i denotes the number which block i is requested at time t. The node prefers to divert the block suffered higher load rather than multi-blocks loaded lighter. The objective of this policy is to reduce the overhead of load diversion and block maintenance overhead. Additionally, block i is not diverted currently.
- The load of target node x is $L_x(t) = \min(L_x(t) < T_a)$, where $x \in M$. Moreover, the target node should locate near to the source node. If there exist several node for diversion, the nearest is selected.

Here, we let $T_t > T_a$ so that the target node leaves more bandwidth for the request from the LHC of itself. Once the load diversion is completed, the source node notifies the clients that are admitted and place a pointer point to target node. If this load is transferred again, the pointer at source node needs to be updated.

2.5 Server Selection from LHC

Here, we adopt the service discovery mechanism presented in Yoid [5]. Firstly, the client leverage some directory system to gain the object's URL. Next, a rendezvous host, IP address of which is resolved by DNS, will return several closest servers to client. Then, the client randomly selects a node as proxy to retrieve the object metadata. Once the requested object metadata is returned, client can start request for the object delivery.

A client requests the $block(i, i = 1, \cdots, N)$ at the beginning and later one succeeding $block(j, j > N)$ per unit time. As the consecutive blocks of an object are mapped randomly, each server maintains the load state of $2N - 1$ units time at most. If only a block is requested at the beginning, each server just maintains the load state of one block time. As a result, it is unable to distinguish between the clients have being severed and the new clients. On the other hand, because a client might cease or FF/RW, subscribing for all blocks at one time will cause unnecessary system load. Moreover, since the streaming holds on a long time, the underlying network condition which fluctuates over time weakens the effect of server selection significantly. We let $N = \mu/T$, where μ is the average reneging time length which is modelled as a normal distribution with $\mu = 5$ minutes [6], and T is time length of block.

As the request message for a block reaches to a leaf node of that LHC from the client's proxy, it will be forwarded to other members by the leaf. The propagation policy of request constrains the the visited range of the request and then impacts the load distribution of LHC. Instead of letting the head of each cluster to forward message to its members at different layers or polling a network center of each cluster to forward the request, we use the constrained flooding for its high robustness, concurrency and minimal delay [1]. Moreover, the flooding method can tolerate the absence of the triangle inequality. Flooded messages can also piggyback a timestamp field to refine the LHC.

Because the request messages might be forwarded by the node to its parent, siblings and children, a few different copies of identical request should be received by same node and only the first one will be forwarded. To distinguish the different copies of identical request message, the node uses Bloom filters to maintain a bit-string which is set and checked by using $hash(client, InitTime, blockID)$. Where, $client$ denotes the client who requests the block, $InitTime$ indicates the time that the request is issued and $blockID$ is the ID of requested block. In order to avoid false positive, this bit-string is resetted at regular intervals or without receipt of any request messages for a long time.

The client also engages in the request forwarding to control the propagation range. As a copy of request arrives at a node with idle resource in the LHC, the node sends a respond message to the client and is pending for echo. The client might receive such a few response messages at the same time and will select a server with minimal IP latency to alleviate the stress of network backbone. In the condition of the local load diversion among a logical server group, the global load balance of blocks placement and two replication polices, the load of server is unnecessary to be a metric in the server selection. After selecting an idle server, the client will notify the other nodes which sent a response to client not to forward message any more.

3 Performance Evaluation

The system performance is restricted by the system size, object popularity and client access behavior. Under the condition of global probabilistic balance, we evaluate the impacts of the local load diversion, server selection, early replication and instantaneous replication policy on the system performance.

3.1 Simulation Model

Total 500 nodes are placed in the plane ranged [0,5000] in the experiment [7]. The storage space of a node is 80GB and the access bandwidth is 100Mb/s. The 2,000 objects are published in the system with four initial replicas. The size of individual object is 3.6GB with length 120 minutes. They are subjected Zipf distribution with skewness factor $\theta = 0.0$. The global request process is modeled as Poison process with parameter $\lambda_G(t)$ [8] determined by the client access behavior:

$$\lambda_G(t) = \frac{N\frac{\lambda}{7}}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2}$$

where, λ is the average number of objects viewed by each client per week and then $\lambda/7$ is the request rate in one day. N denotes the overall number of arrived request in a week. $\mu = 8PM$ is the peak time of arrived request, $\sigma = 45$ minutes is the standard deviation. In addition, we set the local diversion triggering threshold $T_d = 15$, the local diversion acceptance threshold $T_a = 10$, and the early replication threshold $T_{pri} = 5$.

3.2 Resource Utilization

In order to observe the impact of the local load diversion, server selection, early replication and instantaneous replication, we devise the five experiments to observe the contribution of four strategies to the resource utilization of the system. In the test a, all four strategies are used. In addition, the performance loss is observed in absence of the local load diversion, early replication, instantaneous replication and server selection in test b, c, d and e respectively.

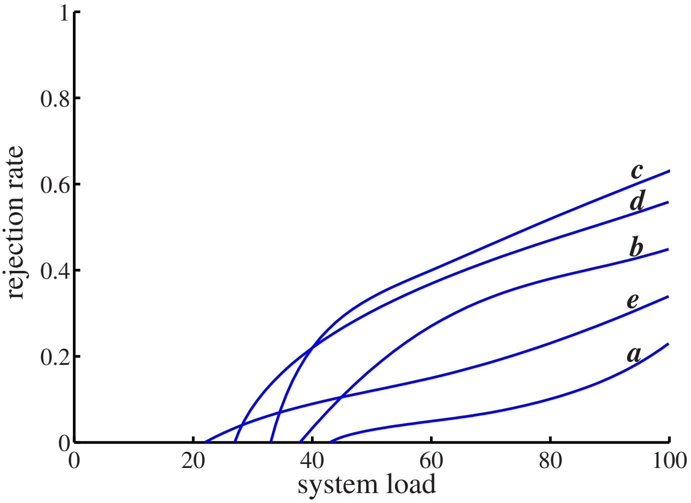


Fig. 3. Rejection rate of various tests schemes

In the figure 3, the rejection rate is low with the light system load showed as curve a; adversely, there is no enough resource and time to diverse load or scale LHC in the high load. In the test b, a client request is refused even there is much available resource in the system due to the load imbalance within a physical cite caused by the probabilistic global load balance and the access skewness of objects. In the scheme represented by curve *c*, the rejection rate rises fast when the system load is high, this is because the instantaneous replication contends a few available resource with objects streaming. In the test *d*, we can observe the rejection rate significantly contributed by the non-popular object under the high system load. Due to the number of clients arrived at various cite is randomly distributed, we intentionally limit the propagation range of the query message in the test *e* as follows: if the first node of LHC visited by the message has no available resource then the instantaneous replication is performed. The result shows that this effect of existing no redirection cannot be eliminated during the whole process although the rejection rate is not high.

4 Conclusions and Future Work

In our scheme, the physical network hierarchy and the logical management hierarchy are substituted by the dynamic LHC for individual block which is expanded with the two different replication policies. As the root of each LHC is mapped into nodes ran-domly, the system is global load balance. We also design the two different replication polices and the local load diversion to improve the utilization further. The logical structure simplifies the server selection largely.

In the future, we will partition each load state into multiple level of sub-state to support the object delivery with various rates.

References

1. Shahabi, C., Banaei-Kashani, F.: Decentralized Resource Management for a Distributed Continuous Media Server. IEEE Transactions on Parallel and Distributed Systems, vol. 13 (2002) 1183–1200
2. Boutaba, R., Hafid, A.: A generic platform for scalable access to multimedia-on-demand systems. IEEE Journal Selected Areas in Communications, Vol. 17 (1999) 1599–1613
3. Cheng-Fu, Chou, Golubchik, L., Lui, J.C.S.: Striping Doesn't Scale: How to Achieve Scalability for Continuous Media Servers with Replication. Int. Conf. on Distributed Computing Systems, (2000) 64–71
4. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. Proceeding IFIP/ACM Middleware (2001)
5. Francis, P.: Yoid: Extending the internet multicast architecture. Unpublished paper, available at http://www.aciri.org/yoid/docs/index.html, (2000)
6. Hua, K.A., Cai, Y., Sheu, S.: Patching: A Multicast Technique for True Video-on-Demand Services. Sixth ACM Int. Conf. Multimedia, (1998) 191–200
7. Zegura, E.W., Calvert, K.L., Bhattacharjee, S.: How to model an internetwork. INFOCOM96, (1996)
8. Nussbaumer, J.-P., Patel, B.V., Schaffa, F.: Multimedia delivery on demand: capacity analysis and implications. The 19th Conf. on Local Computer Networks, (1994) 380–386

Image Forensics Technology for Digital Camera

Jongweon Kim[1], Youngbae Byun[2], and Jonguk Choi[1,2]

[1] College of Computer Software and Media Technology, Sangmyung University
7, Hongji-dong, Jongno-gu, Seoul, 110-743, Korea
{jwkim,juchoi}@smu.ac.kr
[2] MarkAny Inc.
10F, Ssanglim Bldg., 151-11, Ssanglim-dong, Jung-gu, Seoul, 100-400, Korea
byunyb@freechal.com, juchoi@markany.com

Abstract. Even though digital devices, especially digital cameras and camcoders, are getting popular, they are not counted in legal disputes as trustworthy devices. In this paper an image forensics technology is proposed to make digital image capturing devices have legal proof capabilities. For the images taken by digital devices to have legal proof capability, integrity should be guaranteed. Electronic signature was proposed as a means of guaranteeing the integrity of the digital files of images. However, electronic signatures require additional data of digital digest, and cannot survive digital manipulations such as lossy compression, or RST (rotation, scaling, and transformation). This paper suggests a novel algorithm of image forensics that guarantees integrity in nor-mal processing by hiding forensics information into images and identifies locations of forgery and alteration. Images produced by the proposed method maintains high image quality as PSNR over 50[dB] and guarantees integrity up to the quality factor of 85% against JPEG compression.

1 Introduction

With the development of digital technology, analogue devices are being converted into digital ones. Compared to analogue devices, digital devices are so robust against noises and are so easy to manipulate data such as 'store, maintain and edit digital data. Capacity of storage media to store digital data has been continuously expanded with the advancement of semiconductor technology into the direction of larger storage capacity and easier portability.

Even though many digital devices have been introduced in the market and used in everyday life, the raw data captured by and stored into the digital devices cannot be used in legal disputes. In other words, despite advantages of digital devices, as digital data can be easily altered by adversary, the data itself does not have legal proof capabilities. For the reason, computer scientists increasingly pay attention to computer forensics technology [1] that guarantees the integrity of multimedia data.

An effective way to secure legal proofness of collected digital data is digital signature. Digital signature is a mechanism that secures integrity of digital

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 331–339, 2004.

data transmitted between two parties by attaching one-way summarized authentication data, and is becoming a popular in public use with PKI (Public Key Infrastructure). In PKI very commonly the digital signature is conveyed to the receiving party with encrypted text to prove that the original text is not altered during transmission.

However, even though digital signature has been extensively used in text transmission, it cannot be easily applied to proving integrity of multimedia data. First of all, the multimedia data is very frequently altered to enhance efficiency of data transmission and data storage, and thus comparison of received data with original data to check integrity of the data can be meaningless.

This research uses a data hiding technique based on the steganography technology that does not require additional data, as in digital signature mechanism, to prove the integrity of multimedia data. Experimentation has been on images captured by digital camera for the research. To prove integrity of images, hidden information should not be destroyed in compression, but be destroyed in the process of forgery or alteration.

2 Computer Forensics and Image Forensics

The computer forensics collects criminal data that will be used in legal disputes as proofs by certifying that the collected data is authentic [1]. There are two different approaches ensuring integrity of multimedia data by identifying forgery or alteration of the data: digital signature technology [2] and data hiding technology [3].

In order to prove that collected data is authentic and not altered, in PKI the digital signature goes through the following steps. To convey a plain text in the channel, a hash algorithm is applied to generate a hash value, and the hash value is encrypted using a private key, in PKI mechanism, to generate a digital signature. Usually, the digital signature is a hash value encrypted by private key in PKI mechanism and is attached to the plaintext to prove its integrity.

In the digital signature mechanism, modification of a bit of the original data is not practically possible. However, in the process of storing into storage devices or transferring multimedia data through networks multimedia data are frequently compressed. In other words, the size of multimedia data is too large in general, and data compression algorithms are commonly used to reduce the size of multimedia files. Compression of multimedia data is not regarded as forgery or alteration. Thus, applications of digital signature to image data before compression might be meaningless, because of the alteration in the compression processes. In the case of applying digital signature to image data after compression, integrity of the image cannot be verified, because alteration has been done already in the process of compression. Nobody guarantees integrity of the compressed image.

For the reasons, there is a need to develop forensics technology for image data that can utilize the format of multimedia data as it is, and can provide legal proof capability without attaching additional data. One of the technologies that satisfy such requirements is data hiding algorithm that inserts authentication information into digital images to verify its integrity. Data hiding is classified as a kind of steganography that hides information without degenerating the quality of multimedia data. Many algorithms have been suggested, claiming that the algorithm suggested guarantees robustness against various attacks. Low-bit coding method manipulates the lowest bit, while patchwork method uses the statistic characteristics of data by changing the average value of image blocks. Spread spectrum method firstly suggested by Cox et al. [4] is very popular in research community of data hiding uses pseudo random numbers [3].

Usually the data hiding algorithms are classified by the method that inserts information into digital data. However, it can be sometimes classified by according to the degree of the resistance to damage hidden information: robust data hiding and fragile data hiding. In the fragile data hiding, data are easily broken by typical data edition. In the algorithm hidden information is broken even by normal operation of storage and transmission of multimedia data such as lossy compression. Therefore, fragile data hiding algorithms have been suggested to protect hidden information from normal data processing, like lossy compression, and identify data forgery and alteration [5–7]. To implement multimedia forensics, a novel approach is needed to encompass robustness and fragility properties of the data hiding algorithms. Thus, the new approach we suggested in this research is called *semi-fragile data hiding.*

3 Image Forensics Using Semi-fragile Data Hiding

3.1 Linearity of DCT

In DCT used in JPEG compression, the original image is divided into 8×8 pixel blocks and DCT is applied to each block. 2D DCT and IDCT of 8×8 pixels is expressed as follows:

$$\begin{aligned} F(u,v) &= \frac{1}{4}C(u)C(v)\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)\cos\frac{(2x+1)u\pi}{16}\cos\frac{(2y+1)v\pi}{16} \\ f(x,y) &= \frac{1}{4}\sum_{x=0}^{7}\sum_{y=0}^{7} C(u)C(v)F(u,v)\cos\frac{(2x+1)u\pi}{16}\cos\frac{(2y+1)v\pi}{16} \end{aligned} \tag{1}$$

where, $f(x,y)$ is an input image, $F(u,v)$ is the result of transformation (transformed image), and coefficient C is as follows:

$$\begin{aligned} &if\ u=0,\ then\ C(u)=1/\sqrt{2},\ if\ u\neq 0,\ then\ C(u)=1 \\ &if\ v=0,\ then\ C(v)=1/\sqrt{2},\ if\ v\neq 0,\ then\ C(v)=1 \end{aligned} \tag{2}$$

Because DCT and IDCT are summation ($\sum$) as Eq. (1), they satisfy the following equation.

$$af_1(x,y) + bf_2(x,y) \Longleftrightarrow aF_1(u,v) + bF_2(u,v) \tag{3}$$

Thus, DCT has the characteristics of linearity. This research suggests a semi-fragile forensics that embeds authentication information using linearity of DCT.

To increase DCT coefficient (5, 7) by α and (7, 2) by β, multiplication is done for in Eq.(4) by α and $F_1(u, v)$ in Eq.(4) by β, and then summation is done for the two numbers as $F_2(u, v)$ in Eq.(4). Because DCT satisfies linearity as suggested in Eq.(3), $f_3(x, y)$ is the same as Eq.(5). Thus, $f_1(x, y)$ in Eq.(6) is multiplied by α and $f_2(x, y)$ in Eq.(7) multiplied by β. Then they are added to the original data.

That is, using the linear characteristic of DCT, data in DCT domain can be manipulated through only additions and subtractions using data obtained from Eq.(6) and (7) without operations of DCT or IDCT.

$$
\begin{aligned}
&F_3(u,v) = \alpha F_1(u,v) + \beta F_2(u,v)\\
&\begin{vmatrix}
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&\beta&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&\alpha&0&0&0\\
0&0&0&0&0&0&0&0
\end{vmatrix}
= \alpha
\begin{vmatrix}
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0
\end{vmatrix}
+ \beta
\begin{vmatrix}
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0
\end{vmatrix}
\end{aligned}
\tag{4}
$$

$$
f_3(x,y) = \alpha f_1(x,y) + \beta f_2(x,y) \tag{5}
$$

$$
f_1(x,y) =
\begin{vmatrix}
0.068 & -0.068 & -0.068 & 0.068 & 0.068 & -0.068 & -0.068 & 0.068\\
-0.163 & 0.163 & 0.163 & -0.163 & -0.163 & 0.163 & 0.163 & -0.163\\
0.163 & -0.163 & -0.163 & 0.163 & 0.163 & -0.163 & -0.163 & 0.163\\
-0.068 & 0.068 & 0.068 & -0.068 & -0.068 & 0.068 & 0.068 & -0.068\\
-0.068 & 0.068 & 0.068 & -0.068 & -0.068 & 0.068 & 0.068 & -0.068\\
0.163 & -0.163 & -0.163 & 0.163 & 0.163 & -0.163 & -0.163 & 0.163\\
-0.163 & 0.163 & 0.163 & -0.163 & -0.163 & 0.163 & 0.163 & -0.163\\
0.068 & -0.068 & -0.068 & 0.068 & 0.068 & -0.068 & -0.068 & 0.068
\end{vmatrix}
\tag{6}
$$

$$
f_2(x,y) =
\begin{vmatrix}
0.094 & -0.227 & 0.227 & -0.094 & -0.094 & 0.227 & -0.227 & 0.094\\
0.080 & -0.192 & 0.192 & -0.080 & -0.080 & 0.192 & -0.192 & 0.080\\
0.053 & -0.128 & 0.128 & -0.053 & -0.053 & 0.128 & -0.128 & 0.053\\
0.019 & -0.045 & 0.045 & -0.019 & -0.019 & 0.045 & -0.045 & 0.019\\
-0.019 & 0.045 & -0.045 & 0.019 & 0.019 & -0.045 & 0.045 & -0.019\\
-0.053 & 0.128 & -0.128 & 0.053 & 0.053 & -0.128 & 0.128 & -0.053\\
-0.080 & 0.192 & -0.192 & 0.080 & 0.080 & -0.192 & 0.192 & -0.080\\
-0.094 & 0.227 & -0.227 & 0.094 & 0.094 & -0.227 & 0.227 & -0.094
\end{vmatrix}
\tag{7}
$$

3.2 Semi-fragile Data Hiding and Detection

As mentioned earlier, forensics information can be embedded into DCT coefficients of an image only through additions and multiplications using the linearity of DCT. Accordingly, based on the linearity, it is possible selectively to modify

DCT coefficients, to embed semi-fragile forensics information into the spatial domain of the image. For example, there is a Pseudorandom Number (PN) sequence of '1001101'. In the PN sequence code, the left three digits of '100' indicate the abscissa of DCT and the next three digits of '110' indicate the ordinate. The last digit '1' indicates whether the DCT coefficient has been changed or not. Then, '1111011' means that the DCT coefficient of coordinate (4, 6) has been changed. In this case, values in Fig. 1 are added to or subtracted from the spatial domain of the original image. If the DCT coefficient of (4, 6) is to be reduced by 10, the value of Eq.(4) is multiplied by -10 and the result is added to the spatial domain of the original image.

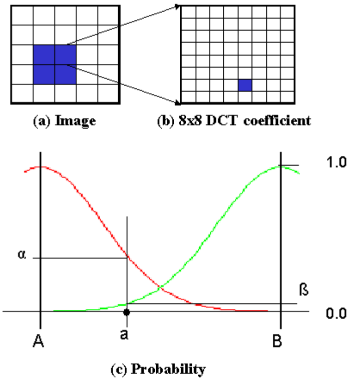


Fig. 1. Method of detecting forgery and alteration

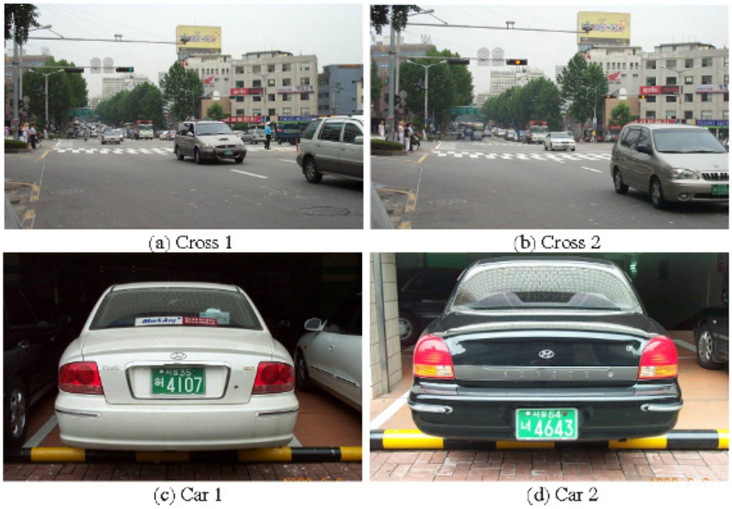
As was done in insertion of forensics information into images in semi-fragile data hiding, the forgery and alteration is detected by determining whether the location of DCT coefficient obtained from PN sequence has been changed or not. However, because the original image is 8-bit gray, errors occur as a result of removing the part of real numbers. In addition, if the image is compressed, errors happen consequently.

To overcome these problems and detect forgeries and alterations, the image is divided into 8 × 8 blocks as in Fig. 1. Then, probability is calculated that the data in each block has been altered or not by checking PN sequence. For example, if the DCT coefficient in (b) of Fig. 1 is α in (c), the probability that the DCT coefficient has not been changed is α and the probability that the DCT coefficient has been changed is β. In this context, there is an equation: $\alpha + \beta = 1$.

Probabilities of each block are calculated and it can be decided 'true' or 'false' by referring PN sequence whether the block has been changed. For example, if PN sequence, which determines the alteration of DCT coefficient, appears to have had been changed, α becomes the value of 'False' and β becomes the value of 'True.' The values of True and False are compared and if the value of True is bigger than that of False it is decided that there is no forgery or alteration in the image, and otherwise it is assumed that there has been alterations.

4 Experiments and Discussion

The images taken by digital cameras are can be used as a proof of criminal evidence by legal enforcement organizations. However, because digital images are very easily manipulated by simple operations with graphic tools and editing programs, they cur-rently do not have legal proof capability in legal disputes. In this research, taken were a number of photographs of traffic situations and parked cars to experiment effective-ness of semi-fragile algorithms. Illegal parking and traffic sign violation is very hot social issues in Korea, and forgery and alteration of digital images of the scenes draw extensive attention. The digital camera used in the experiment was Kodak DC280 of 2.8 million pixels. This research tested alteration of traffic signs and number plates, and checked the PSNR of photographs, in which information of integrity verification was inserted to confirm that the photographs are not different from their originals.


(a) Cross 1 (b) Cross 2

(c) Car 1 (d) Car 2

Fig. 2. Test images

Images used in the experiment were four 800 × 600 size images as in Fig. 2. When forensics information was added to the four original images, the PSNR (Peak Signal to Noise Ratio) was measured as in Table 1. As in Table 1, the image forensics technology used in this paper shows high PSNR over 50[dB] on the average, even after forensics information was added to the original images. This indicates that, although forensics information was added, the images can be used without feeling of degeneration because the damage rate of the original images is quite low.

As mentioned earlier, a certain degree of robustness against lossy compression like JPEG is necessary in image forensics. In general, JPEG lossy compression

Table 1. PSNR results of the test images

Image	Cross1	Cross2	Car1	Car2	Mean
PSNR[dB]	50.41	50.33	50.42	50.28	50.36

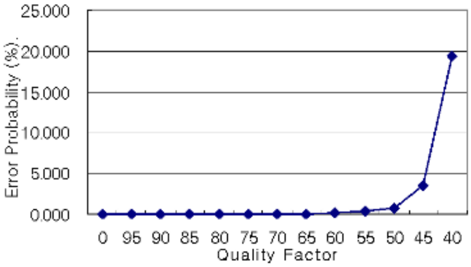


Fig. 3. Error probabilities for JPEG compression

uses an image quality factor to compress. If an image is compressed at a quality factor of 90%, the file size is reduced down to less than one tenth. To confirm the forensics performance of the technology used in this paper, error probability according to the quality factor of JPEG lossy compression is presented in Fig. 3.

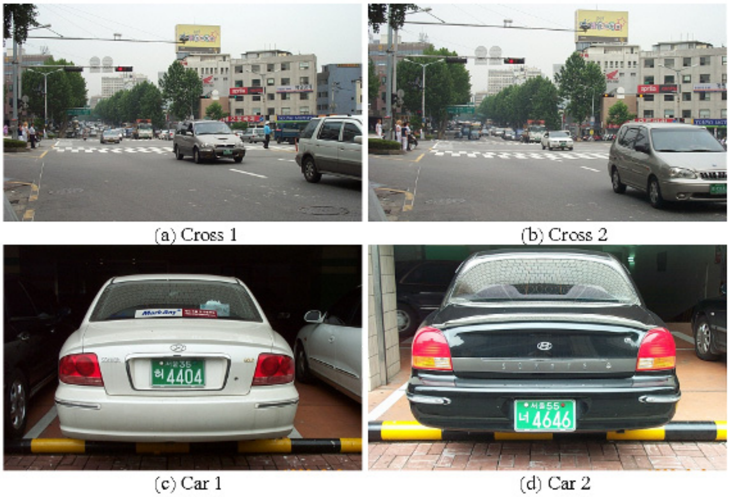
As shown in Fig. 3, the image forensics technology suggested in this paper has error probability of zero in JPEG lossy compression at a quality factor of up to 85%. Thus, it can prove the absence of forgery and alteration. In addition, its error prob-ability is less than 1% at a quality factor of 50%, and thus the technology is effective in determining forgery and alteration. Because the general quality factor used by digital cameras is 90 95% the proposed technology can be applied usefully.

Figure 4 shows images resulting from altering the images given in Fig. 4 using an image editor. In the two intersection images, the traffic signs, which had been green and yellow respectively, were altered red. In the two car images, the numbers were altered from 4107 to 4404, from 54 to 55 and from 4643 to 4646.

Figure 5 shows the result of forensics detection on the altered images in Fig. 4. Figure 5 shows that the traffic signs and number plates in the digital camera images have been altered using an image editor.

5 Conclusion

As optical cameras are replaced with digital ones and image editors are growing more sophisticated with the development of computer technologies, the legal proof capability of digital images is questioned. It is quite difficult for a non-expert to forge and alter images from optical film, but digital images are alterable by anybody who has a computer so they are not considered as legal evidence.



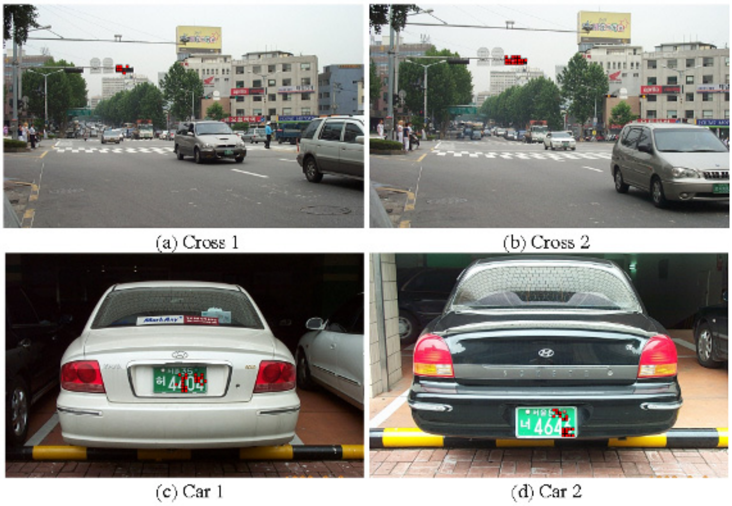
(a) Cross 1

(b) Cross 2

(c) Car 1

(d) Car 2

Fig. 4. Altered images



(a) Cross 1

(b) Cross 2

(c) Car 1

(d) Car 2

Fig. 5. Detection of the altered area

This paper proposes an image forensics technology that can prove integrity of digital camera images by embedding forensics information simultaneously when the images are taken by camera. The proposed method gained PSNR of over 50[dB], a degree at which no difference from the original image is detected. In particular, it was confirmed that the technology guarantees integrity without loss for a quality factor of up to 85% in JPEG compression, which is a popular lossy compression algorithm for the storage and transmission of images. In the digital signature technology, which is frequently used in proving the integrity of digital data, digital signature has to be regenerated in JPEG compression and forgeries and alterations by one who can generate an electronic signature are not detectable. Thus the image forensics technology proposed in this study is expected to be quite effective.

As discussed above, digital signature can guarantee the integrity of digital data, but image data must be accompanied with an additional digital signature. Furthermore, the technology cannot be used with normal processes such as JPEG compression or it requires the generation of a new electronic signature whenever such processes are executed. Especially because the image owner can alter the image and attach a new digital signature the technology cannot prevent forgeries and alteration by the owner.

The proposed technology guarantees the integrity of digital camera images or digital camcorder records, so makes them admissible as legal evidence. In addition, it minimizes the use of computing resources using the linearity of DCT, so reduces additional costs for digital cameras or digital camcorders to the minimum.

Acknowledgements. Solideo Gloria. This work was supported by National Research Laboratory project(2000N-NL-01-C-286) of Ministry of Science and Technology in Korea.

References

1. W. G. Kruse II and J. G. Heiser, Computer Forensics, Addison-Wesley, Boston, 2001
2. http://www.itl.nist.gov/div897/pubs/fip186.htm
3. W. Bender, D Gruhl, N. Morimoto, and A. Lu, "Techniques for Data Hiding", IBM Systems Journal, Vol. **25**, pp. 313–335, 1996
4. I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia", NEC Res. Inst., Princeton, NJ, Tech. Rep. 95–10, 1995
5. C.-Y. Lin and S.-F. Chang, "Semi-Fragile Watermarking for Authenticating JPEG Visual Content", Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, pp. 140–151, January 2000
6. Kurato Maeno, et al, "New Semi-Fragile Image Authentication Watermarking Techniques Using Random Bias and Non-Uniform Quantization", Proc. SPIE Security and Watermarking of Multimedia Contents, San Jose, California, pp. 659–670, January 2002
7. Eugene T. Lin, et al. "Detection of image alterations using semi-fragile watermarks", Proc. SPIE Security and Watermarking of Multimedia Contents, San Jose, California, pp. 23–28, January 2000

Lossless Data Hiding Based on Histogram Modification of Difference Images*

Sang-Kwang Lee[1], Young-Ho Suh[1], and Yo-Sung Ho[2]

[1] Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong, Yuseong-gu, Deajeon, 305-350, Korea
{sklee,syh}@etri.re.kr
[2] Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea
hoyo@gist.ac.kr

Abstract. In this paper, we propose a new lossless data hiding method where distortion due to data embedding can be completely removed from the watermarked image after the watermark has been extracted. In the proposed method, we utilize characteristics of the difference image and modify pixel values slightly to embed the data. We show that the lower bound of the PSNR (peak-signal-to-noise-ratio) values for typical images are about 51.14 dB. Moreover, the proposed method is quite simple and fast. Experimental results demonstrate that the proposed scheme can embed a large amount of data while keeping high visual quality of test images.

Keywords: Lossless data hiding, watermarking, histogram modification

1 Introduction

Digital representation of multimedia content offers various advantages, such as easy and wide distribution of multiple and perfect replications of the original content. However, the fact that an unlimited number of perfect copies can be illegally produced is a serious threat to the right of content owners. In order to protect the intellectual property rights, we can apply information hiding techniques in various application areas, such as broadcast monitoring, proof of ownership, content authentication, copy control, and transaction tracking [1].

In most data hiding techniques, the original image is inevitably distorted due to data embedding itself. Typically, this distortion cannot be removed completely due to quantization, bit replacement, or truncation at the gray level 0 and 255. Although the distortion is often quite small, it may be unacceptable for medical or legal imagery or images with a high strategic importance in certain military

* This work was supported in part by Gwangju Institute of Science and Technology (GIST), in part by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC) at GIST, and in part by the Ministry of Education (MOE) through the Brain Korea 21 (BK21) project.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 340–347, 2004.

applications [2]. Thus, it is desired to reverse the watermarked image back to the original image after the embedded data are extracted. Data embedding satisfying this requirement, is referred to as lossless data hiding.

In recent years, several lossless data hiding techniques have been proposed for images. Lossless data embedding can take place in the spatial domain [2,3,4], or in the transform domain [5,6]. Ni et al. [4] proposed a lossless data embedding technique, which utilizes the zero or the minimum point of the image histogram. It can embed a large amount of data and the PSNR values of watermarked images are always higher than 48 dB. However, gray level values of the zero point and the peak point should be transmitted to the receiving side for data retrieval.

In this paper, we propose a new lossless data hiding method where we exploit the difference image histogram to embed more data than other lossless data hiding schemes. The proposed scheme gives about 3 dB improvement in PSNR for typical images as compared to Ni's scheme [4]. Moreover, there is no need to transmit any side information to the receiving side for data retrieval.

This paper is organized as follows. In Section 2, we describe details of the proposed lossless data hiding method using the histogram modification of the difference image. After experimental results are presented in Section 3, we conclude this paper in Section 4.

2 Proposed Lossless Data Hiding Scheme

2.1 Watermark Embedding

Figure 1 shows the watermark embedding procedure of the proposed scheme, which consists of watermark generation, creating the difference image, histogram shifting, and histogram modification.

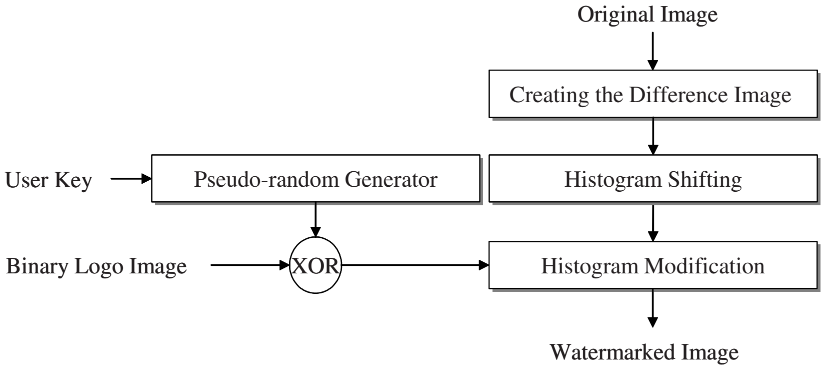


Fig. 1. Proposed watermark embedding

In order to generate a binary watermark sequence $W(m,n)$ of size $P \times Q$, we combine a binary random sequence generated by the user key, $A(l)$ of length $P \times Q$ with a binary logo sequence $B(m,n)$ of size $P \times Q$ pixels using the bit-wise XOR operation.

$$W(m,n) = A(l) \oplus B(m,n), \quad 0 \le l \le P \times Q - 1, \quad 0 \le m \le P-1,\ 0 \le n \le Q-1 \tag{1}$$

For a grayscale image $I(i,j)$ of size $M \times N$ pixels, we form the difference image $D(i,j)$ of size $M \times N/2$ from the original image.

$$D(i,j) = I(i,2j+1) - I(i,2j),\ 0 \le i \le M-1,\ 0 \le j \le \frac{N}{2} - 1 \tag{2}$$

where $I(i,2j+1)$ and $I(i,2j)$ are the odd-line field and the even-line field, respectively. For watermark embedding, we empty the histogram bins of -2 and 2 by shifting some pixel values in the difference image. If the difference value is greater than or equal to 2, we add one to the odd-line pixel. If the difference value is less than or equal to -2, we subtract one from the the odd-line pixel. Then, the modified difference image $\widetilde{D}(i,j)$ can be represented as

$$\widetilde{D}(i,j) = \widetilde{I}(i,2j+1) - I(i,2j) \tag{3}$$

where

$$\widetilde{I}(i,2j+1) = \begin{cases} I(i,2j+1)+1 & \text{if } D(i,j) \ge 2 \\ I(i,2j+1)-1 & \text{if } D(i,j) \le -2 \\ I(i,2j+1) & \text{otherwise} \end{cases} \tag{4}$$

In the histogram modification process, the watermark $W(m,n)$ is embedded into the modified difference image $\widetilde{D}(i,j)$. The modified difference image is scanned. Once a pixel with the difference value of -1 or 1 is encountered, we check the watermark to be embedded. If the bit to be embedded is 1, we move the difference value of -1 to -2 by subtracting one from the odd-line pixel or 1 to 2 by adding one to the odd-line pixel. If the bit to be embedded is 0, we skip the pixel of the difference image until a pixel with the difference value -1 or 1 is encountered. In this case, there is no change in the histogram. Therefore, the watermarked fields $I_w(i,2j+1)$ and $I_w(i,2j)$ are obtained by

$$I_w(i,2j+1) = \begin{cases} \widetilde{I}(i,2j+1)+1 & \text{if } \widetilde{D}(i,j) = 1 \text{ and } W(m,n) = 1 \\ \widetilde{I}(i,2j+1)-1 & \text{if } \widetilde{D}(i,j) = -1 \text{ and } W(m,n) = 1 \\ \widetilde{I}(i,2j+1) & \text{otherwise} \end{cases} \tag{5}$$

and

$$I_w(i,2j) = I(i,2j) \tag{6}$$

2.2 Watermark Extraction and Recovery

Figure 2 depicts the watermark extraction and recovery procedure. In this process, we extract the binary logo image and reverse the watermarked image to the original image.

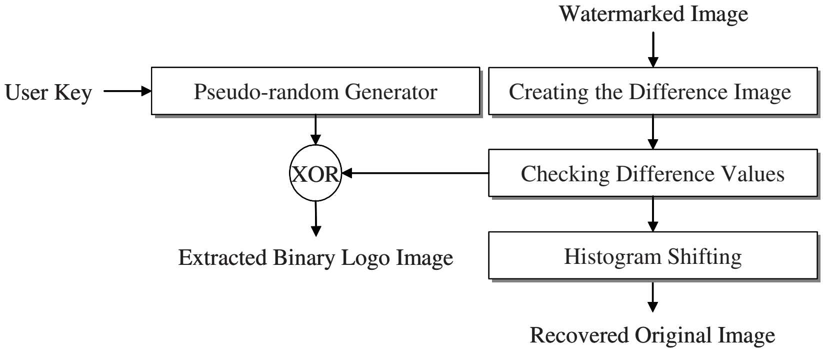


Fig. 2. Proposed watermark extraction and recovery

We calculate the difference image $D_e(i,j)$ from the received watermarked image $I_e(i,j)$. The whole difference image is scanned. If the pixel with the difference value of -1 or 1 is encountered, the bit 0 is retrieved. If the pixel with the difference value of -2 or 2 is encountered, the bit 1 is retrieved. In this way, the embedded watermark $W_e(m,n)$ can be extracted.

$$W_e(m,n) = \begin{cases} 0 & \text{if } D_e(i,j) = -1 \text{ or } 1 \\ 1 & \text{if } D_e(i,j) = -2 \text{ or } 2 \end{cases} \tag{7}$$

In order to reconstruct the binary logo image $B_e(m,n)$, we perform the bitwise XOR operation between the binary random sequence generated by the user key, $A_e(l)$ and the detected binary watermark sequence $W_e(m,n)$.

$$B_e(m,n) = A_e(l) \oplus W_e(m,n) \tag{8}$$

Finally, we reverse the watermarked image back to the original image by shifting some pixel values in the difference image. The whole difference image is scanned once again. If the difference value is less than or equal to -2, we add one to the odd-line pixel. If the difference value is greater than or equal to 2, we subtract one from the odd-line pixel. The recovered odd-line field $I_r(i,2j+1)$ can be expressed as

$$I_r(i,2j+1) = \begin{cases} I_e(i,2j+1) - 1 & \text{if } D_e(i,j) \geq 2 \\ I_e(i,2j+1) + 1 & \text{if } D_e(i,j) \leq -2 \\ I_e(i,2j+1) & \text{otherwise} \end{cases} \tag{9}$$

Since we manipulate pixel values of only the odd-line field in the watermark embedding process, the recovered even-line field $I_r(i,2j)$ is

$$I_r(i,2j) = I_e(i,2j) \tag{10}$$

2.3 Lossless Image Recovery

The proposed scheme cannot be completely reversed because the loss of information occurs during addition and subtraction at the boundaries of the grayscale range (at the gray level 0 and 255). In order to prevent this problem, we adopt modulo arithmetic for watermark addition and subtraction. For the odd-line field $I(i, 2j+1)$, we define the addition modulo c as

$$I(i, 2j+1) +_c 1 = (I(i, 2j+1) + 1) \bmod c \tag{11}$$

where c is the cycle length. The subtraction modulo c is defined as

$$I(i, 2j+1) -_c 1 = (I(i, 2j+1) - 1) \bmod c \tag{12}$$

The reversibility problem arises from pixels with truncated due to overflow or underflow. Therefore, we use $+_c$ and $-_c$ instead of + and - only when truncation due to overflow or underflow occurs. In other words, we have only to consider $255 +_c 1$ and $0 -_c 1$.

In the receiving side, it is necessary to distinguish between the cases when, for example, $I_e(i, 2j+1) = 255$ was obtained as $I(i, 2j+1)+1$ and $I(i, 2j+1) -_{256} 1$. We assume that no abrupt change between two adjacent pixels occurs. If there is a significant difference between $I_e(i, 2j+1)$ and $I_e(i, 2j)$, we estimate that $I(i, 2j+1)$ was manipulated by modulo arithmetic.

$$\begin{cases} I(i, 2j+1) + 1 & \text{if } |I_e(i, 2j+1) - I_e(i, 2j)| \le \tau \\ I(i, 2j+1) -_{256} 1 & \text{otherwise} \end{cases} \tag{13}$$

where τ is a threshold value. Similarly, $I_e(i, 2j+1) = 0$ is estimated as

$$\begin{cases} I(i, 2j+1) - 1 & \text{if } |I_e(i, 2j+1) - I_e(i, 2j)| \le \tau \\ I(i, 2j+1) +_{256} 1 & \text{otherwise} \end{cases} \tag{14}$$

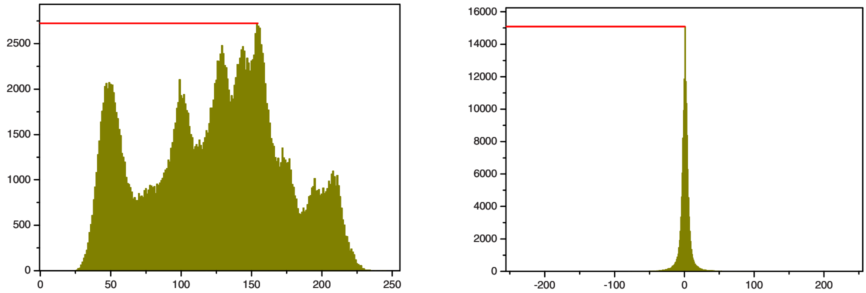
2.4 Lower Bound of PSNR and Embedding Capacity

Assume that there is no pixel with overflow and underflow in the original image. In the worst case, all pixels of the odd-line field will be added or subtracted by 1. The MSE (mean squared error) of this case is 1/2. Hence, the PSNR of the watermarked image can be calculated as

$$\mathrm{PSNR(dB)} = 10 \log_{10}(255^2 \cdot 2) \approx 51.14 \tag{15}$$

In short, the lower bound of the PSNR of the watermarked image is about 51.14 dB. This result is much higher than other lossless data hiding techniques.

The embedding capacity of this scheme equals to the number of pixels with the difference values of -1 and 1 in the difference image. A large number of pixel values of the difference image have a tendency to be distributed around 0. Using this property of the difference image histogram, we can embed a large amount of



(a) Histogram of the original image

(b) Histogram of the difference image

Fig. 3. Histogram Characteristics of Lena image

data as compared to the original image itself. Figure 3 shows this characteristic property of difference images. The number of pixels with the peak point in the histogram of the original Lena image is around 2,750. On the other hand, the number of pixels with the peak point in the histogram of the difference image is higher than 15,000.

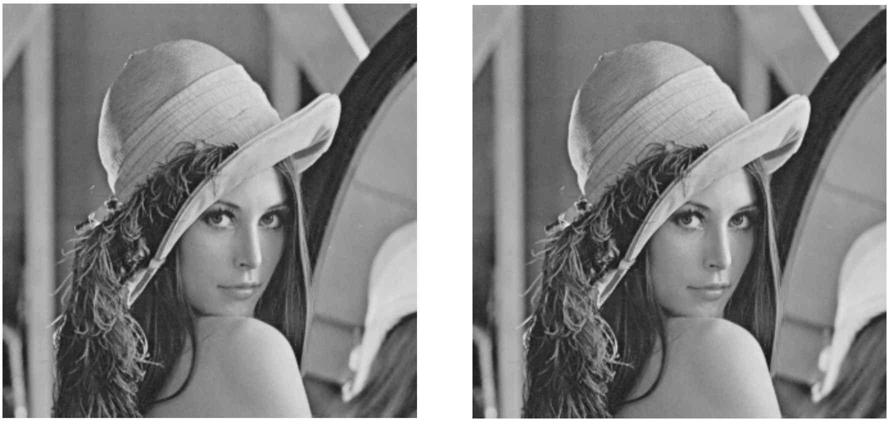
3 Experimental Results and Analysis

In order to evaluate the performance of the proposed scheme, we perform computer simulations on many typical grayscale images of size 512 × 512 pixels. Figure 4 shows a watermark which is a binary logo image of size 128 × 56 pixels, equivalent to a binary sequence of 7,168 bits.

Fig. 4. Binary logo image of 128 × 56 pixels

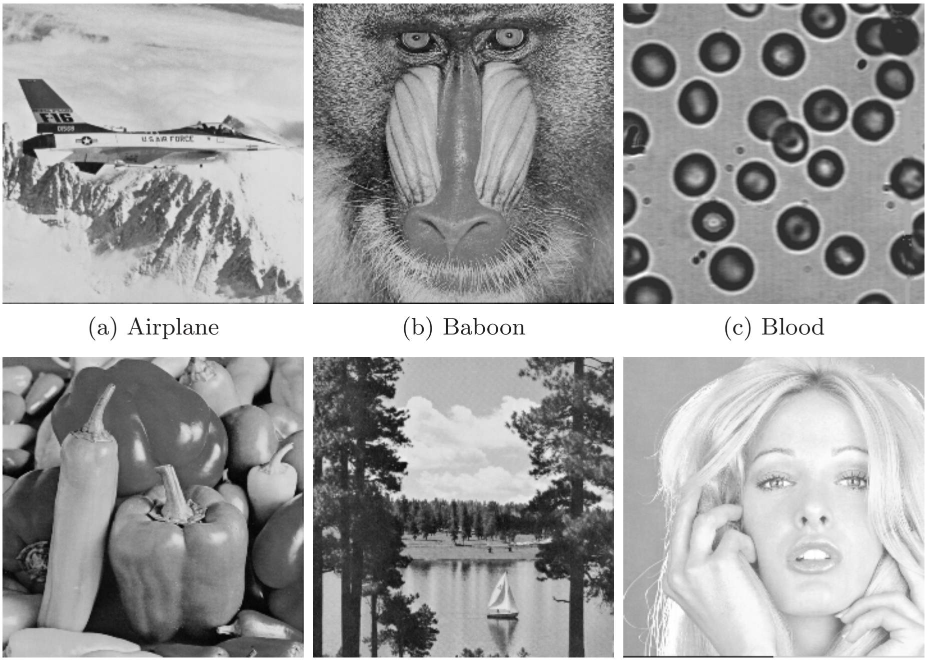
The original and watermarked Lena images are shown in Fig. 5. The Lena image does not contain pixels with truncated due to overflow or underflow. It is observed that there is no visible degradation due to embedding in the watermarked image. Figure 6 shows further six watermarked images.

Table 1 summarizes the experimental results. This table shows that the PSNR values of all watermarked images are above 51.14 dB, as we theoretically proved in Section 2.4. The capacity ranges from 8 kbits to 30 kbits for 512 × 512 × 8 test grayscale images. This result shows that the proposed scheme offers adequate capacity to address most applications. It is also seen from Table 1 that an image like Baboon, which contains significant texture, has considerably lower capacity than simple images such as Airplane.



(a) Original image (b) Watermarked image

Fig. 5. Results with Lena image



(a) Airplane (b) Baboon (c) Blood

(d) Peppers (e) Sailboat (f) Tiffany

Fig. 6. Watermarked images

Table 1. Experimental results

Test images (512×512×8)	PSNR (dB)	Capacity (bits)	Overflow/underflow (No. of pixels)
Airplane	58.78	30,487	0
Baboon	51.49	7,383	0
Blood	55.59	22,009	20
Lena	57.63	23,579	0
Peppers	55.74	17,280	2
Sailboat	55.55	14,391	0
Tiffany	52.50	26,004	83

Some test images, such as Blood, Peppers, and Tiffany, contain pixels with overflow and underflow. However, the set of such pixels is relatively small and the artifacts due to overflow and underflow are not serious.

For simplicity, we used different values of -1 and 1 for watermark embedding. If we use the different value of 0 instead of -1 or 1, we can ensure larger capacity than that shown in the experiment.

4 Conclusions

We have proposed a lossless data hiding method based on difference image histogram. In order to solve the reversibility problem, we used the modulo arithmetic instead of the ordinary addition and subtraction. Experimental results showed that the proposed scheme provides high embedding capacity while keeping the embedding distortion as small as possible. Reversibility back to the original content is highly desired in sensitive imagery, such as military data and medical data. The proposed lossless data hiding technique can be deployed for such applications.

References

1. Lee, J., Hwang, S., Jeong, S., Yoon, K., Park, C., Ryou, J.: A DRM framework for distributing digital contents through the Internet. ETRI Journal (2003) 423–436
2. Fridrich, J., Goldjan, M., Du, R.: Invertible authentication. Proc. SPIE, Security and Watermarking of Multimedia Contents (2001) 197–208
3. Honsinger, C., Jone, P., Rabbani, M., Stoffel, J.: Lossless recovery of an original image containing embedded data. US Patent: 6,278,791 B1 (2001)
4. Ni, Z., Shi, Y., Ansari, N., Su, W.: Reversible data hiding. Proc. ISCAS (2003) 912–915
5. Goldjan, M., Fridrich, J., Du, R.: Distortion-free data embedding. Proc. 4th Information Hiding Workshop (2001) 27–41
6. Xuan, G., Zhu, J., Chen, J., Shi, Y., Ni, Z., Su, W.: Distortionless data hiding based on interger wavelet transform. IEE Electrionics Letters (2002) 1646–1648

A Robust Image Watermarking Scheme to Geometrical Attacks for Embedment of Multibit Information

Jung-Soo Lee[1,2] and Whoi-Yul Kim[2]

[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 proposes a new watermarking scheme that is robust to geometrical attacks and can embed large user information. By inserting a template that is used for correcting geometrical distortions we can embed user information into an image affordably. We embed the user information using a base watermark (128x128 pixel size). If the base watermark is shift then the peak position of its cross-correlation with the watermarked image is also shift. Using this fact, our algorithm can embed large information. To extract user information we have to firstly extract the template. After restoring the watermarked image to the original state using the template, we can extract the embedded user information. The performance of the presented scheme is evaluated for various geometrical attacks and compression. Results show that the fact that the scheme is robust to geometrical distortions and effective to embed and extract large user information.

1 Introduction

During the last decade, digital watermarking technologies have been developed to protect the ownership of digital contents. Because of the various attacking methods, however, finding out the technique that is applied for all attacks is very difficult [1,2,3]. In particular, the geometrical attacks change the positions of the image pixels, which make the watermarking system difficult to detect the watermark. Furthermore, efforts to deal with geometrical attacks prevent the capacity of information to be embedded from increasing [4,5,6].

To solve these problems, we propose the watermarking system for providing the robustness to geometrical attacks and for increasing the capacity of the information to be embedded.

Previous Work
Kutter proposed a periodic watermark insertion method for obtaining the information of geometrical affine transform [4]. The periodic embedment of water-

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 348–355, 2004.

mark plays an important role in extracting the geometrically transformed information of the watermarked image. It makes peaks spring up periodically through an autocorrelation of a watermarked image. But as a general rule, because the peaks are weakly sprung, it is too difficult to distinguish them correctly.

Pereia and Pun proposed a method that embeds a template with watermark into an image to deal with geometrical attacks [5]. 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 [6]. 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.

This Paper is composed of four sections. In section 2, we explain the method that embed watermark into an image. Next the process of watermark extracting is described in section 3. To make synchronization between an input image and base watermark matched, we extract the template firstly. And then the method for extracting the user information is explained. In section 4, experimental results are showed. And we conclude about our study in section 5.

2 Watermark Embedding

We embed two kinds of watermark into an image. One is user information and the other is template. User information is embedded in the spatial domain using shift control. By shifting the base watermark according to the user information, we can control the peak position of the cross-correlation between the base watermark and the watermarked image. And the template is embedded in the frequency domain and helps extracting the exact user information by recovering the watermarked image to the original state (geometrical distortions free).

2.1 Embedding User Information

The procedure to embed user information into an image is depicted in Fig. 1.

The input user information (text or Arabian numeral) is firstly converted to binary sequences. Using the base watermark that is random data composed of {-1, 1}, we convert binary sequences into the information watermark(W_{II}). As shifting the base watermark according to the binary sequences(each 4bits) and adding the shift base watermarks, we are able to make the information watermark(W_{II}).

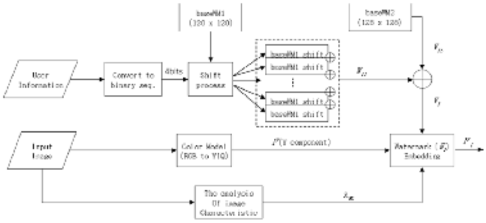


Fig. 1. The embedment of the user information.

The shift control is performed using eq. 1. That is to say, we have to shift the base WM1 to the extent of S_x horizontally and S_y vertically according to 4bits' binary sequences.

$$S_x = \left\{\mathrm{mod}_{(B_M/B_S)}(I_U)\right\} \times B_s + \frac{B_s}{2} + \left\{\mathrm{mod}_{(B_L/B_M)}(I_{U-nth} - 1)\right\} \times B_M$$
$$S_y = \left\lfloor \frac{I_U}{(B_M/B_S)} \right\rfloor \times B_S + \frac{B_S}{2} + \left\lfloor \frac{(I_{U-nth} - 1)}{(B_L/B_M)} \right\rfloor \times B_M \qquad (1)$$

Where, I_U ($\in$\{0,1,$\cdots$, 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. mod_p(x) means the remainder resulting from dividing x into p. I_{U-nth} (1^{st}, $\cdots$, 16^{th}) means where 4bits' data to be embedded locate(n th 4bits). For example, if $B_L = 128$, $B_M = 32$, $B_S = 8$, $I_{U-nth} = 9$ and $I_U = 6$ ('0110' expressed in binary code), we should shift the base WM1 to 20 pixels horizontally and 76 pixels vertically. To embed 64bits into the image with base WM1, we have to repeat the shift and adding process 16 times. W_{IS} is another base watermark that has a different seed value from base WM1. It is to cope with the cropping attack. Because our watermarking scheme extracts the user information using the peak position of cross-correlation baseWM1 with a watermarked image, if cropping attack is not corrected, we cannot extract the right user information.

To embed information watermark(W_I) into an image, we use the following equation.

$$I^y_I = I^y + \lambda_{wc} W_I \qquad (2)$$

Where, I^y_I indicates the watermarked image where information watermark is embedded. I_y indicates the Y(luminance) component of the input image. And λ_{wc} regulates the strength of W_I . λ_{wc} is calculated through the following process.
a. Apply high pass filter to the Y component using the high pass filter mask.
b. Normalize the output above to an appropriate level.

2.2 Embedding Template

To cope with geometrical distortions, we embed template into I^y_I for obtaining the geometrically deformed information. Our template differs from that of Pereia

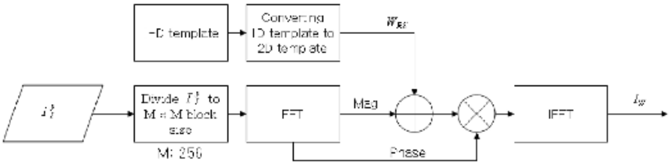

1-D template
Converting 1D template to 2D template
W_{RS}
I_I^y
Divide I_I^y to M × M block size
M: 256
FFT
Mag
Phase
IFFT
I_W

Fig. 2. The embedding scheme of W_{RS}.

and Pun[5]. While they insert the 8-point peak, we successively insert random sequences into middle frequencies of the input image. Because our method does not have to find local peaks, it is able to extract readily the information of geometrical attacks. That is, if the position where random sequences are inserted is found using the correlation, we can easily know the rotation and scaling information using its sequential data. The detail explanation of these contents is addressed in Subsection 3.1. The embedding process of the template is followed below.

2-D template (W_{RS}) is made through process such as Fig. 3.

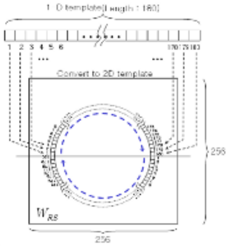


Fig. 3. Composing the 2-D template (W_{RS}).

After generating the 1-D random sequence composed of {-1, 1}, we successively align it roundly in empty 2-D space sizing 256x256 such as Fig. 3. Because the magnitude component of Fourier transform has the origin of symmetry, we symmetrically align 1-D random sequence like as the lower part of W_{RS}.

The generated W_{RS} is added to the magnitude component of I_I^y in DFT(Discrete Fourier Transform) domain using the Eq. 3.

$$I_W^M = F_M\{I_I^y\} + \mu_{WC} \times W_{RS} \tag{3}$$

Here, I_W^Mindicates the result of addition the spectrum(magnitude component) of the input image to W_{RS}. And $F_M\{I_I^y\}$ is the spectrum of the input image. And μ_{WC} is a constant for controlling the strength of W_{RS}. After the result above is combined with the phase component of DFT of I_I^y, the watermarked image(I_W) is made through the inverse DFT. And I_W is converted to RGB model combining with I and Q components that have been saved in the embedding process of information watermark.

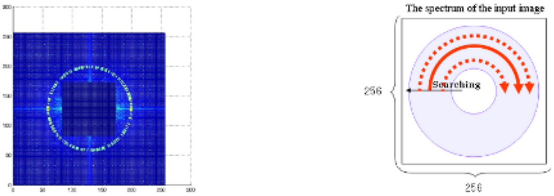
3 Watermark Extracting

In the watermark extraction process, we have to firstly extract the template to descry the geometrically deformed information of the watermarked image. The obtained 2-D template is converted to 1-D template through circular projection. After correcting the watermarked image using the extracted information from the template, we extract the user information from the restored watermarked image.

3.1 The Extraction of Template

Figure 4 describes the template extraction process.

Firstly, we divide an input image into 256 by 256 size. Next FFT is performed on the block image(256 by 256). If the input image has been watermarked, its spectrum is shown like as (a) in Fig. 4 (to meet the convenience of sight, we remove low frequencies). (b) explains the process of extracting one dimensional template from the spectrum of the input image. The belt(where W_{RS} is added) shown in fig. 4 (a) is moved when the watermarked image is resized. This belt, which has 1-D template used in the embedding process of W_{RS}, is found through cross-correlation the original 1-D template with the roundly extracted 1-D sequential data from the spectrum of the input image.



(a) the spectrum of the watermarked image (b) searching 1-D template

Fig. 4. The spectrum of the watermarked image and the searching process.

Once the belt is found, we extract the 1-D template. If rotation is taken place in the watermarked image, we can get the rotation angle by finding the position of cross-correlation peak, which is obtain from correlation between the original 1-D template and the extracted 1-D template, such as Fig. 5.

And if scaling was occurred in it, we could know the degree of scaling by finding distance from the origin to the position of the belt in the spectrum of the input image spectrum like as Fig. 6.

3.2 Extraction of User Information

Using the extracted 1-D template, we solved the problem about the rotation and the scaling distortions. And the problem of translation can be solved by

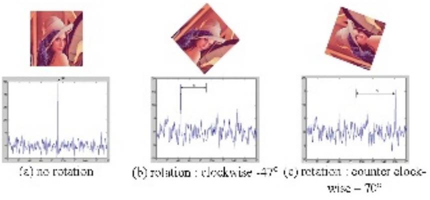

(a) no rotation (b) rotation : clockwise -47° (c) rotation : counter clock-wise – 70°

Fig. 5. Cross-correlation between 1-D template and extracted 1-D template when rotation is occurred to the watermarked image.

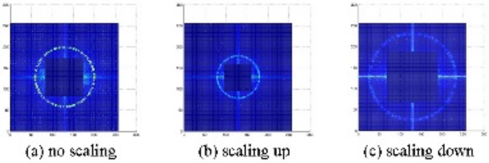

(a) no scaling (b) scaling up (c) scaling down

Fig. 6. The shape of the spectrum of the watermarked image when scaling is occurred.

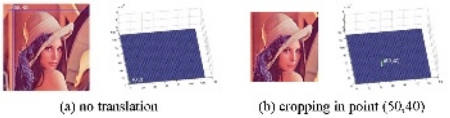

(a) no translation (b) cropping in point (50,40)

Fig. 7. Cross-correlation between W_{IS} and the watermark image when cropping or shift is occurred in it.

using W_{IS}. In fig. 7, cross-correlation between the watermarked image and W_{IS} is displayed.

If the watermarked image is corrected to the original state, we can extract user information by finding cross-correlation between the watermarked image and baseWM1.

Equation 4 explains the scheme that extracts the embedded information.

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

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

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. I_{u-nth} means the position where the extracted 4bits' data are located.

4 Experimental Results

In this section, we test our scheme to ascertain robustness against the geometrical attacks, JPEG compression, and other various attacks.

We apply various attacks to the watermarked image and examine the bit error rate from the extracted information. In the limited extent, proposed approach successfully recovered the embedded information. In JPEG compression test, QF means the image quality factor of the target image.

Table 1. Test results about geometrical distortions.

Distortions	Proposed approach		Pereira's approach	
	BER(%)	Message length	*BER*(%)	Message length
Rotation (1°~359°)	0	80 bits	0	72 bits
Scaling (50% < S < 200%)	5.6		33	
Cropping (> 96 × 96 pixels)	0		21	
JPEG comp. (> QF 40%)	0		35.7	
Blurring (> PSNR 38dB)	0		12	
Color depth reduction (> 4bits/pixel)	0		0	
Sharpening	0		0	
Noise addition(< 50%)	3.8		40	

In table 1, *BER* means bit error rate that is expressed in following formula.

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

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

5 Conclusions

In this paper, we proposed a new watermarking scheme that is robust to geometrical attacks and embeds large bits of information. To find out the inserted template in template detection process, we do not choose the local point peaks like as [5] but successively calculate the position where the template is embedded through the correlation between the roundly extracted data(like as fig. 4 (b)) and the original 1-D template. Using the extracted template, we can restore the watermarked image from the distorted status, which makes it possible to extract the user information. We also embedded the user information using the shift control. We verified that the shift control method for embedding the user information is more efficient to extract it than the method that extracts it using the magnitude of correlation, which has to choose the threshold for the watermark detection. Besides the shift control method has the characteristic that can embed more large information using the same sized base watermark.

Acknowledgements. 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. Kutter.: Watermarking resisting to translation, rotation and scaling. Proc. SPIE, vol. 3528, Nov. (1998), 423–431
5. 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
6. 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)

Combined Encryption and Watermarking Approaches for Scalable Multimedia Coding

Feng-Cheng Chang, Hsiang-Cheh Huang, and Hsueh-Ming Hang

Department of Electronics Engineering, National Chiao Tung University,
Hsinchu 300, Taiwan, R.O.C.,
hmhang@mail.nctu.edu.tw

Abstract. Intellectual Property (IP) protection is a critical element in a multimedia transmission system. Conventional IP protection schemes can be categorized into two major branches: *encryption* and *watermarking*. In this paper, a structure to perform layered access protection by combining encryption and robust watermarking is proposed and implemented. By taking advantage of the nature of cryptographic schemes and digital watermarking, the copyright of multimedia contents can be well protected. We adopt the scalable transmission method over the broadcasting environment. The embedded watermark can be thus extracted with high confidence. Then, the next-layer secrets can be perfectly decrypted and reconstructed. Finally, the media contents are recovered.

1 Introduction

With the widespread use of multimedia broadcasting, the digital media, including images, audio and video clips, are easily acquired in our daily life. The current network environments make scalable coding of multimedia a necessary requirement when multiple users try to access the same information through different communication links [1,2]. Scalability means that a multimedia data bitstream is partitioned into layers in such a way that the base layer is independently decodable into a content with a reduced quality. The reduction may be in spatial resolution, temporal resolution, or signal-to-noise ratio (SNR). To reproduce the original content, enhancement layers provide additional data to restore the original quality from the base layer. Enhancement layers represent the scalability of the content coding, namely, spatial, temporal, or SNR scalability. Therefore, scalable coding of multimedia is suitable for delivering digital contents to different users and devices with various capabilities [3].

In many cases, it requires to deliver multimedia content securely. However, the channel for multimedia broadcasting is an open environment; thus, if the user data and information are not protected, it might be illegally used and altered by hackers To protect privacy and intellectual property (IP) rights, people often use cryptographic techniques to encrypt data, and the contents protected by encryption are expected to be securely transmitted over the Internet [4,5].

In cryptography, the contents to be encrypted are called *plaintext*, and the encrypted contents are called *ciphertext*. Although cryptographic schemes provide secure data exchange among peers, it implies that the ciphertext cannot be

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 356–363, 2004.

altered during transmission [6]. If any one bit is received erroneously, the plaintext cannot be decrypted correctly. This is not a good property when we deliver protected contents in a broadcasting environment, where erroneous transmission may occur occasionally. There, a one-bit error may cause a totally useless content. To meet this deficiency for multimedia broadcasting, we include watermarking technique to aid encryption, because the watermarked contents can withstand some kind of *attacks*, including signal processing, geometric distortion, and transmission errors. In this paper, we combine both the cryptographic and watermarking techniques for layered content protection. On the one hand, the message for protection of multimedia contents can be perfectly decrypted by cryptography, while on the other hand, the encrypted message can be further protected by robust watermarking algorithms to resist transmission errors.

This paper is organized as follows. Sec. 2 describes the concepts and issues of layered content protection. In Sec. 3, we propose a layered protection structure with combined cryptographic and watermarking schemes. We give an application example and simulations in Sec. 4. And Sec. 5 concludes this paper.

2 Layered Protection Concepts

As discussed in Sec. 1, scalable coding is a solution to broadcast contents to devices with various playback capabilities. With the nature of layered coding, the whole media can be partitioned into blocks of data. Thus, it is straightforward to group receivers of different playback capabilities by sending different combinations of data partitions. However, the conditional access (CA) requirement is dealt in a different way in a broadcast environment. To distinguish different groups of users, a popular solution is to encrypt data by a group-shared key. Thus, the CA issue can be solved by encrypting data partitions with different keys, and a granted user has the corresponding decryption keys to the assigned data partitions.

The next issue is how to distribute the keys. Depending on the delivery infrastructure, two problems may arise. One is how to protect keys from malicious listeners. There are methods to protect keys from malicious listeners, such as the one proposed in the DVB standard [7]. The other problem is how to synchronize (in time) a key with the content. For example, to broadcast a protected content over Internet, we may send the key to users via a reliable channel (such as RTSP connection [8]), while the content goes through an unreliable channel (such as RTP sessions [9]). A reliable channel guarantees information correctness by sacrificing delivery speed, and it is likely that the key information is out-of-sync to the corresponding content.

A possible solution to eliminate synchronization problem is to transmit the key information together with the content, such as inserting it into the optional header fields of the coded stream. However, it may be destroyed by transmission errors or transcoding. Our proposed method is less sensitive to minor transmission errors. We embed the key information into the content with robust watermarking techniques. Since the key information is available at the same time as we reconstruct the content, the (time) synchronization problem is resolved. The drawback of this approach is that if packet loss or transcoding occurs, the reconstructed content is different from the original one, and the key information may not be extracted accurately. To reduce the impact of unreliable or distorted delivery, we incorporate robust digital watermarking methods [10] to reinforce the robustness of the embedded key information.

The main steps of the layered protection is organizing secrets (keys and necessary parameters) into a watermark, robustly watermarking the base layer, and encrypting the enhancement layer. A granted user receives the base layer, extracts and derives the decryption key, decrypts the enhancement layer, and combine layers together to produce the contents. In the following sections, we will describe our proposed method in detail.

3 Proposed Method

In this section, we describe the layered decryption and decoding operations on the receiver side. Because the associated encryption and encoding operations vary depending on the scalable coding, we provide an example at the end of this section. We first describe the receiver architecture in our proposed method, then we describe the corresponding transmitter architecture in the following paragraphs.

3.1 Receiver Architecture

Scalable coding is composed of one *base layer* and several *enhancement layers* to match the network diversity for transmission. The enhancement operation is illustrated in Fig. 1. Assuming that the initial base layer B_0 has been received, the subsequent composing operations can be expressed by

$$B_i = \text{compose}\ (B_{i-1}, E_i), \tag{1}$$

where

$$E_i = \text{decrypt}_e\ (X_i, K_i). \tag{2}$$

In Eq. (1), B_{i-1} is the available base layer, and E_i is the enhancement layer to improve quality from B_{i-1} to B_i. During transmission, E_i is protected by a cryptic algorithm with K_i as the key, and the transmitted data is X_i in Eq. (2).

There are some secret information to be obtained prior to decrypting E_i, and the operations can be expressed as follows:

$$W_i = \text{extract}\ (B_{i-1}, P_{i-1}) \tag{3}$$

$$F_i = \text{decrypt}_f\ (W_i, G_i) \tag{4}$$

$$K_i = \text{key}\ (F_i) \tag{5}$$

$$P_i = \text{param}\ (F_i) \tag{6}$$

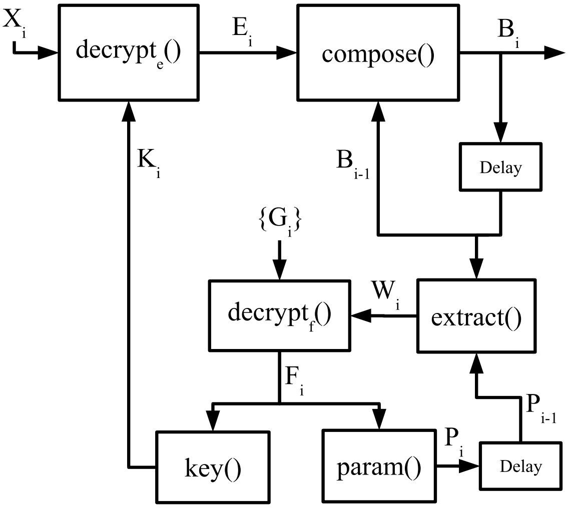


Fig. 1. Decryption and decoding of layer-protected content

W_i is the digital watermark extracted from the constructed base layer B_{i-1} with extraction parameter P_{i-1}. As described in Sec. 2, W_i represents the protected secret information. Thus, we have the secret information F_i by decrypting the watermark using user-specific key G_i. After parsing F_i, we obtain the decryption key K_i and the next watermark extraction parameter P_i.

As Fig. 1 illustrates, the decryption and composition blocks are iterative processes. There are several initial parameters required to activate these processes. We will discuss how to obtain the initial parameters in the following paragraphs.

- When the whole content is protected, namely, B_0 is encrypted, we need K_0 to decrypt X_0. In this case, K_0 should be obtained by a separate channel.
- One scenario is that B_0 is the "preview" layer; i.e., B_0 is not encrypted, we simply bypass the decrypt$_e$ block.
- Depending on the watermarking algorithm, the extraction process may requires specific parameters. If it does, the first watermark extraction parameter P_0 should be obtained in a separate channel to activate subsequent extraction process.
- All the key-decryption keys $\{G_i\}$ should be obtained before receiving the media data, for instance, by manually or automatically update after subscription.

3.2 Transmitter Architecture

Depending on the scalable coding algorithm, the design of transmitter side varies. Fig. 2 shows one of the possible designs. The architecture is almost the inverse of the receiver architecture in Fig. 1. The watermark W_i is the encrypted version of the key K_i and the embedding parameter P_i. The B'_{i-1} is the un-watermarked

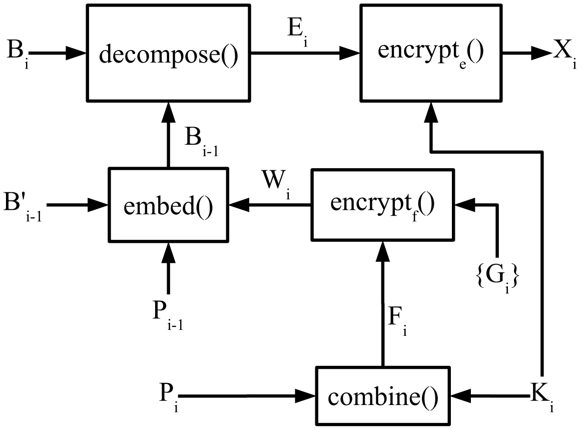


Fig. 2. Encryption and encoding of layer-protected content

base layer with lower quality. After embedding W_i into B'_{i-1}, we have the base layer B_{i-1}. The enhancement layers are generated as the differences between B_i and B_{i-1}. All the $\{K_i\}$, $\{P_i\}$, and $\{G_i\}$ are known in advance.

4 Simulation Results

In this paper, we use the test image `Lena` with size 1024 × 1024 to conduct the simulations in this section. The original `Lena` is first converted to 512 × 512 base layer. The DES[11] key (8 ASCII letters **“NCTU-DEE”** in Fig. 3(a)) to encrypt enhancement layer is also encrypted using DES by the user key ($\{G_i\}$ in Fig. 1) to generate the 8-byte (or 64-bit) secret. The secret is then repeated for 32 times to form the binary watermark, as shown in Fig. 3(b).

Fig. 3. Plaintext encryption and watermark generation. (a) The 8-byte plain-text. (b) The converted binary watermark with size 128 × 128.

Figure 4 shows the data in transmitted base layer and the enhancement layers. Before transmission, the watermarked base layer has acceptable visual quality, with the PSNR of 39.24 dB in Fig. 4(a). We then extract the watermark from the base layer picture, derive the decryption key, decrypt the transmitted enhancement data in the next layer, and finally reconstruct the original 1024 × 1024 picture.

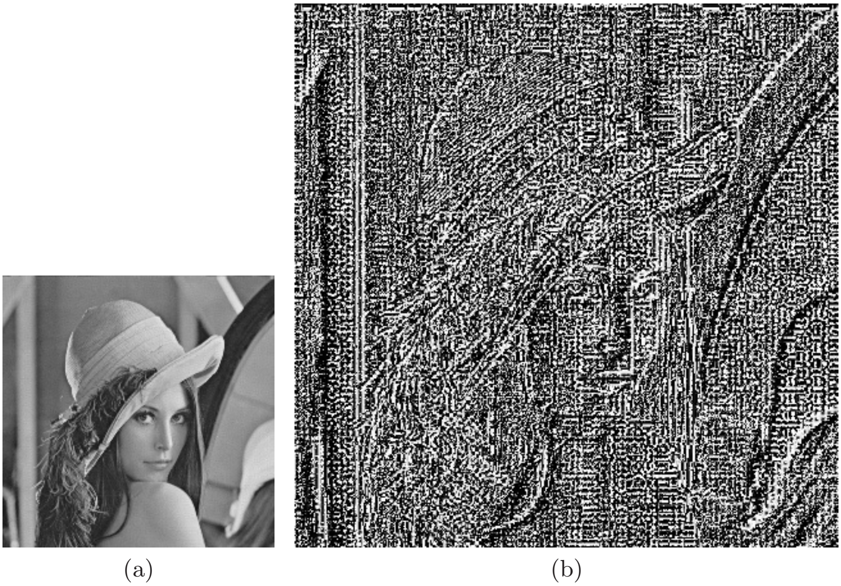


Fig. 4. (a) 512 × 512 base layer. (b) 1024 × 1024 enhancement layer.

We then test the packet loss case on the base layer [12]. The packet loss rate in our simulations is set to 10%. The extracted watermark is shown in Fig. 5(a). The distortion is within the tolerance range of the extracted watermark, with the bit-correct rate of 92.74%. We then use the majority vote to produce the 8-byte secret, extracted encryption key, and decrypt the ciphertext. Finally, we can recover the original key information correctly as shown in Fig. 5(b). In addition, the 1024 × 1024 picture thus can be reconstructed with some defects as shown in Fig. 6.

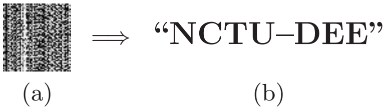


Fig. 5. Watermark extraction and cipher-text decryption. (a) The extracted watermark, with the bit-correct rate of 92.74%. (b) The decrypted cipher-text, which is identical to that in Fig. 3(a).

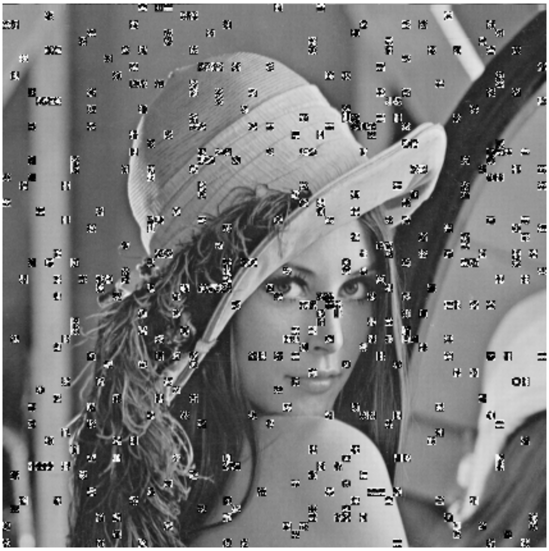


Fig. 6. The encrypted and watermarked image corrupted by transmission errors, with best-effort reconstruction.

5 Conclusion

In this paper, we proposed a structure to protect the layered (scalable) content in a broadcast environment. By combining cryptographic and robust watermarking techniques, the secret for decrypting enhancement data streams can be safely embedded in the base layer. Robust watermark enables embedding information directly in the multimedia content, and the embedded bits can be extracted even when the watermarked media experience attacks during transmission. On the other hand, cryptography provides confidentiality. But it does not tolerate any bit error. The contribution in this paper is to combine these two techniques, and offer the advantages of both for intellectual property protection.

In the proposed scheme, the encryption concept guarantees the access control, keeping away malicious eavesdroppers. Also, the embedding concept solves the key-content synchronization problem. The robust watermarking concept increases the data robustness against transmission errors and distortions. Comparing to conventional cipher-block chaining encryption, our method not only provides a way to guarantee access controls, but also synchronously transmits decryption information. Moreover, robust watermarking implicitly gives higher data integrity protection on the keys than on the contents. One simulated example demonstrates the effectiveness of the proposed structure.

In our future work, we will modify our structure with scalable video coding. We will also integrate our proposed structure with the MPEG IPMP (Intellectual Property Management and Protection) message exchange format[13,14].

References

1. Sun, X., Wu, F., Li, S., Gao, W., Zhang, Y.-Q.: Seamless switching of scalable video bitstreams for efficient streaming. IEEE Transactions on Multimedia **6** (2004) 291–303
2. Almeida, J.M., Eager, D.L., Vernon, M.K., Wright, S.J.: Minimizing delivery cost in scalable streaming content distribution systems. IEEE Transactions on Multimedia **6** (2004) 356–365
3. Wiegand, T., Sullivan, G.J., Bjntegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. IEEE Transactions on Circuits and Systems for Video Technology **13** (2003) 560–576
4. Parviainen, R., Parnes, P.: Large scale distributed watermarking of multicast media through encryption. Proceedings of the International Federation for Information Processing, Communications and Multimedia Security Joint Working Conference IFIP TC6 and TC11 (2001) 149–158
5. Lim, Y., Xu, C., Feng, D.D.: Web-based image authentication using invisible fragile watermark. Conferences in Research and Practice in Information Technology (2002) 31–34
6. Xu, X., Dexter, S., Eskicioglu, A.M.: A hybrid scheme for encryption and watermarking. IS&T/SPIE Symposium on Electronic Imaging 2004, Security, Steganography, and Watermarking of Multimedia Contents VI Conference (2004) 723–734
7. Digital video broadcasting project (DVB): http://www.dvb.org/ (2004)
8. Real time streaming protocol: http://www.rtsp.org/ (2004)
9. Schulzrinne, H.: http://www.cs.columbia.edu/~hgs/rtp/ (2004)
10. Shieh, C.S., Huang, H.C., Wang, F.H., Pan, J.S.: Genetic watermarking based on transform domain techniques. Pattern Recognition **37** (2004) 555–565
11. Data Encryption Standard (DES): http://www.itl.nist.gov/fipspubs/fip46-2.htm (1993)
12. Chande, V., and Farvardin, N.: Progressive transmission of images over memoryless noisy channels. IEEE Journal on Selected Areas in Communications **18** (2000) 850–860
13. Avaro, O., Eleftheriadis, A., Herpel, C., Rump, N., Swaminathan, V., Zamora, J., Kim, M.: MPEG systems (1-2-4-7) FAQ, version 17.0. ISO/IEC JTC1/SC29/WG11 N4291 (2001)
14. Huang, C.C., Hang, H.M., Huang, H.C.: MPEG IPMP standards and implementation. IEEE PCM'02 (2002) 344–352

Digital Image Watermarking Using Independent Component Analysis

Viet Thang Nguyen and Jagdish Chandra Patra

School of Computer Engineering, Nanyang Technological University,
Nanyang Avenue, Singapore 639798
thangnguyen@pmail.ntu.edu.sg, aspatra@ntu.edu.sg

Abstract. In many watermarking applications, tracking the image copy ID is another requirement besides the ownership verification. Furthermore, it is a common demand to keep the original image from public accessibility during the watermark extraction process. In this paper, we propose a new watermarking method called WMica that bases on the Independent Component Analysis (ICA) technique. The proposed method employs a two-watermark embedding scheme; one watermark is to identify the ownership and the other serves as the ID for each copy of the original image. In the extraction scheme, an ICA algorithm is applied together with a down-sizing technique so that we can estimate all the watermarks without accessing the original image and the prior information about the watermarks. The new method, undergoing a variety of experiments, has shown its robustness against many salient attacks. It also exhibits a capability in image authentication.

1 Introduction

Digital Watermarking, in which some information is embedded directly and imperceptibly into digital data to form watermarked data, is one of the most effective techniques to protect digital works from piracy and has been being extensively studied recently [1]. To estimate the watermark, some algorithms require the original image to be available in the extraction. It is not favored since the owners always prefer to hide their original works from public accessibility. Moreover, in various applications, it is important to mark each copy with an unique number, i.e, the image copy ID, so that the authors can track these copies. It means the watermarking method developed for these applications should not require the watermark information in the extraction.

ICA is an important technique in signal processing field for estimating unknown signals from their observed mixtures [2]. With its blind extraction capability, several researchers have been trying to employ ICA in watermarking. The authors in [3, 4] propose algorithms that partition off the original image and the watermark into independent components (ICs), then combine these ICs to produce the watermarked image. This technique, however, requires a lot of computation and usually fails in brute-force attacks. Another approach considers the original image as one unknown source signal, and the watermark is another

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 364–371, 2004.

unknown signal. The watermarked image is then a mixture of these signals. This approach is simple to implement but usually need additional knowledge about the original data or the watermark. For example, in [5, 6], the algorithm needs both the secret key image and the original image.

In this work, we developed a novel watermarking method called WMica that employs the second ICA-based approach above. However, we further exploit the characteristics of the image to overcome the need of original image and big-sized key image. We do the extraction on the down-sized version of the test image, thus we only have to use a small-sized key image instead of the original image. A down-sizing and up-sizing methods has been applied to create all the mixtures for ICA algorithm only from the test image and this small-sized key image. Moreover, instead of a single watermark, the algorithm embeds two watermarks into the host image, one for identifying the ownership and one for identifying the copy ID. Comparing with other watermarking techniques, our proposed method has the following advantages: (i) The ICA-based extraction scheme does not need the original image and knowledge of the watermark. (ii) The 'key image' is a small-sized image and can be publicly available. (iii) The two watermark system help the owners both to verify their ownership and to track the image's copying. (iv) The proposed watermarking algorithm can serve as both robust watermarking and fragile watermarking.

2 The Proposed WMica Method

The most important task of WMica is to create enough mixture signals for ICA extraction only from the test image and a small-sized key image. To do it, we employ down-sizing and up-sizing functions and some special modifications on the watermarks in the embedding scheme. Upon the watermark extraction scheme, the watermarked image (test image) is down-sized to different sizes and combined with the key image to create the mixtures. The mixtures then serve as the input for ICA algorithm for estimating the watermarks. The details on these functions are provided in the following paragraphs.

Denote by $I_{M\times N}$ the image of size $M \times N$, the two functions up-sizing $\mathcal{U}(I_{M\times N}, k)$ and down-sizing $\mathcal{D}(I_{M\times N}, k)$ are defined as follows

$$\begin{aligned} \mathcal{D}(I_{M\times N}, k) &= I_{[k]\frac{M}{k}\times\frac{N}{k}} \qquad & I_{[k](m,n)} &= \tfrac{1}{k^2}\textstyle\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} I_{(km+i,kn+j)} \\ \mathcal{U}(I_{M\times N}, k) &= I^{[k]}_{kM\times kN} \qquad & I^{[k]}_{(km+i,kn+j)} &= I_{m,n} \end{aligned} \tag{1}$$

where k is a positive integer called 'resizing factor'. $I_{[k]\frac{M}{k}\times\frac{N}{k}}$ and $I^{[k]}_{kM\times kN}$ are defined as the k times down-sized image and k times up-sized image of $I_{M\times N}$, respectively. $i, j = 0, 1, ..., k-1$, $m = 0, 1, ..., M-1$ and $n = 0, 1, ..., N-1$. The value of each pixel $I_{(m,n)}$ is ranged between $[-1, 1]$. In the next step, the two watermarks W_1 and W_2 are modified from the owner's signatures to satisfy[1]

$$\mathcal{D}(W_1, k_1k_2) = \emptyset \tag{2}$$

[1] From this point, the image size index $_{M\times N}$ in the term $I_{M\times N}$ is omitted when it is not necessary.

$$\mathcal{D}(W_2, k_1) = \mathcal{U}(\mathcal{D}(W_2, k_1 k_2), k_2) \tag{3}$$

where $\emptyset_{M\times N}$ denotes an image whose pixel values are all zero. Then we have

$$\mathcal{D}(aW_1 + bW_2, k_1) = a\mathcal{D}(W_1, k_1) + b\mathcal{D}(W_2, k_1) \tag{4}$$

$$\mathcal{U}(\mathcal{D}(aW_1 + bW_2, k_1 k_2), k_2) = a\emptyset + b\mathcal{D}(W_2, k_1). \tag{5}$$

Equation (4) and (5) clearly represent two mixtures that are produced from the only signal $aW_1 + bW_2$. Combining with the signal taken from the key image, we have enough three mixtures for ICA extraction.

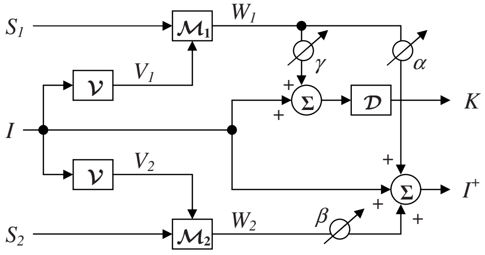


Fig. 1. The proposed embedding scheme.

2.1 The Embedding Scheme

A scheme of the embedding process is displayed in Fig. 1. From the initial signatures S_1 and S_2 (generally are the small images), two watermarks W_1 and W_2 are created through the modification functions $\mathcal{M}_1(\cdot)$ and $\mathcal{M}_2(\cdot)$ so that these two watermarks satisfy (2) and (3), respectively. In addition, visual masks V_1 and V_2 are applied into $\mathcal{M}_1(\cdot)$ and $\mathcal{M}_2(\cdot)$. With the help of visual mask, one can increase the watermark strength considerably while maintaining the image quality and the watermark invisibility. In this paper, we apply a Noise Visibility Function (NVF) [7] as the function $\mathcal{V}(I, L)$ to create visual mask. Pixels in a visual mask V are computed from the image I by

$$V_{(m,n)} = \frac{1}{1 + \sigma_I^2(m, n)} \tag{6}$$

where $\sigma_I^2(m,n)$ denotes the local variance of the image in a window centered on the pixel $I_{(m,n)}$. $\sigma_I^2(m,n) = \frac{1}{(2L+1)^2}\sum_{i=-L}^{L}\sum_{j=-L}^{L}(I_{(i+m,j+n)} - \bar{I}_{(m,n)})^2$ with $\bar{I}_{(m,n)} = \frac{1}{(2L+1)^2}\sum_{i=-L}^{L}\sum_{j=-L}^{L} I_{(i+m,j+n)}$ where a window of size $(2L+1)\times(2L+1)$ is used for the estimation.

Next, the two watermarks W_1 and W_2 are inserted into the original image I to get I^+, the watermarked image. Mean while, the key image K is acquired by down-sizing a combination of I and the first watermark W_1. The embedding steps involved can be summarized as follows

1. Create two visual masks V_1 and V_2 with the window length L_i $(i = 1, 2)$

$$V_1 = \mathcal{V}(I, L_1); \qquad V_2 = \mathcal{V}(I, L_2). \tag{7}$$

2. Create two watermarks W_1 and W_2 that satisfy (2) and (3)

$$W_1 = \mathcal{M}_1(S_1, V_1, k_1, k_2); \qquad W_2 = \mathcal{M}_2(S_2, V_2, k_1, k_2). \tag{8}$$

3. Create the watermarked image I^+ and the key image K

$$I^+ = I + \alpha W_1 + \beta W_2 \tag{9}$$

$$K = \mathcal{D}(I + \gamma W_1, k_1). \tag{10}$$

Parameters α and β are called ‘embedding coefficients’, γ is called ‘keying coefficient’. These parameters can be any non-zero values in the range of $[-1, 1]$.

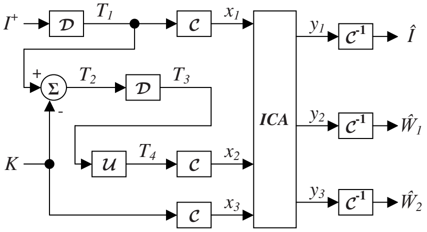


Fig. 2. The ICA-based extraction scheme.

2.2 The Watermark Extraction Scheme

A scheme of the ICA-based extraction is shown in Fig. 2. Watermarks are estimated from the image I^+ with the help of key image K which is assumed to be accessible during extraction. The ICA-based extraction steps include:

1. Down-size the watermarked image I^+ to the size of the key image K

$$T_1 = \mathcal{D}(I^+, k_1) \tag{11}$$

2. Create the image T_4 from T_1 and K with up-sizing and down-sizing methods

$$T_4 = \mathcal{U}(\mathcal{D}(T_1 - K, k_2), k_2). \tag{12}$$

3. Create one-dimensional signals from T_1, T_4 and K

$$[x_1, x_2, x_3]^T = [\mathcal{C}(T_1), \mathcal{C}(T_4), \mathcal{C}(K)]^T \tag{13}$$

   where $\mathcal{C}(\cdot)$ denotes the 2-to-1 dimension converter.
4. Apply ICA technique on $\mathbf{x} = [x_1, x_2, x_3]^T$ to get three outputs $\mathbf{y} = [y_1, y_2, y_3]^T$.
5. Convert back the outputs $\mathbf{y}$ to images

$$(\hat{I}, \hat{W}_1, \hat{W}_2) = \mathcal{C}^{-1}(\mathbf{y}) \tag{14}$$

   where $\mathcal{C}^{-1}(\cdot)$ is a 1-to-2 dimension converter. $\hat{I}$, $\hat{W}_1$ and $\hat{W}_2$ denote the estimates of the down-sized version of the original image I and the two watermarks W_1 and W_2, respectively.

3 Experiments and Results

The experiments were implemented on the gray-scale images Lena and Baboon of size 512 × 512. The embedded watermarks were created from two small images of size 64 × 64, including an author signature and a copy ID. To measure and control the quality of the watermarked image, Peak Signal to Noise Ratio ($PSNR$) was selected as a criterion. To ensure the high quality of the watermarked image and the imperceptibility of the watermarks, the embedding coefficients were chosen so that $PSNR > 42dB$ in all experiments. The values of these parameters and the other setting are provided in Table 1.

Table 1. Configuration of the experiments.

	α	β	γ	k_1	k_2	L	$PSNR$
Expt.1	$\frac{9}{256}$	$\frac{11}{256}$	$-\frac{6}{256}$	4	2	15	46.25
Expt.2	$\frac{9}{256}$	$\frac{11}{256}$	$-\frac{6}{256}$	4	2	15	42.84

The inputs and outputs of the embedding scheme are shown in Fig. 3 (please not that all the negative values in the watermarks were plotted by their absolute values so that they can be displayed in the figures).

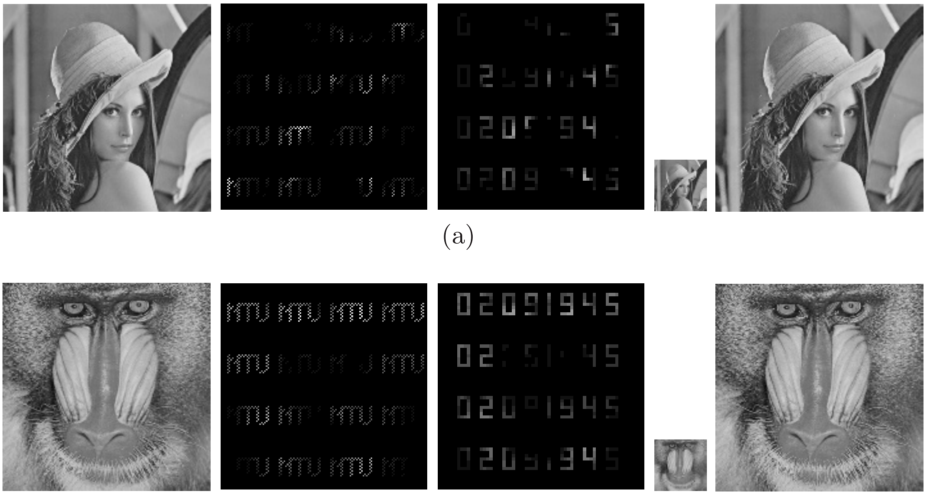
To assess the quality of the estimates, we select the correlation coefficient r between the down-sized version of the watermark and its estimate. The correlation coefficient r between two images $X_{M\times N}$ and $Y_{M\times N}$ is defined as

$$r = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{(i,j)} - \bar{X})(Y_{(i,j)} - \bar{Y})}{\sqrt{(\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{(i,j)} - \bar{X})^2)(\sum_{i=1}^{M}\sum_{j=1}^{N}(Y_{(i,j)} - \bar{Y})^2)}} \tag{15}$$

where $\bar{X} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} X_{(i,j)}$ and $\bar{Y} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{(i,j)}$.

3.1 JPEG Compression and Gray Level Reduction Test

The watermarked image I^+ was compressed using JPEG with different quality factors before the watermarks were extracted. The correlation coefficient between an estimate $\hat{W}_i$ and its original watermark $W_{i[k_1]}$ $(i = 1, 2)$ was computed for each quality factor and plotted in Fig. 4(a). The proposed algorithm provided very good performance on all experiments. The qualities of the estimates are high even when the JPEG compression quality factor is reduced awfully. Even in the lowest quality setting (quality factor = 10%), most of the watermarks are still recognizable ($r > 0.4$).

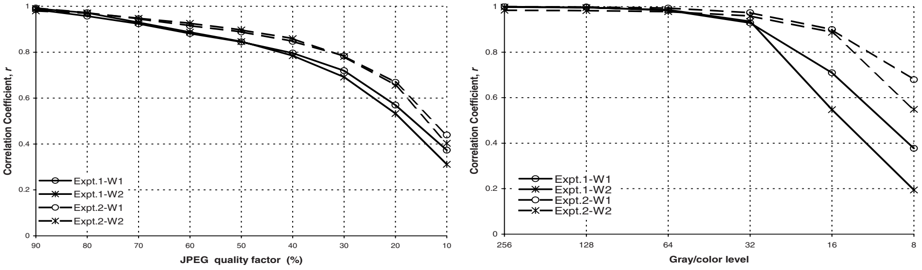


(a)

(b)

Fig. 3. From left to right: Original image I, Watermarks W_1 and W_2, Key image K and Watermarked image I^+. (a) Expt 1. (b) Expt 2.

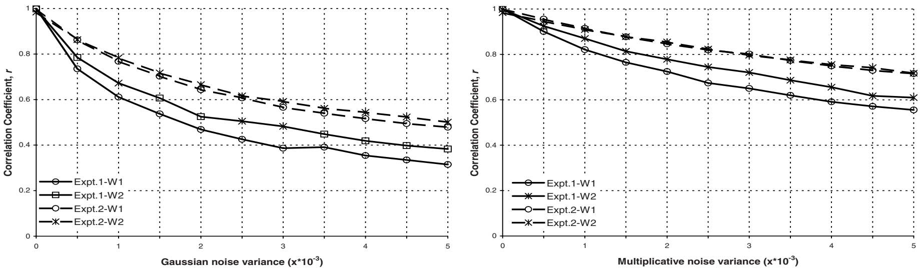
The algorithm offered excellent results in the next test, the gray level reduction. In this modification, the gray level of the watermarked image I^+ was reduced from 256 level down to $128, 64, ..., 8$ level. The values of the performance index r (shown in Fig. 4(b)) in all experiments are close to each other and the estimated watermarks are highly correlated to the original ones. It can be seen that the proposed algorithm is able to extract the signature successfully ($r > 0.4$) when the gray scale level is reduced up to a level of 16.



(a)

(b)

Fig. 4. Results of (a) JPEG compression test and (b) gray level reduction test.



(a) (b)

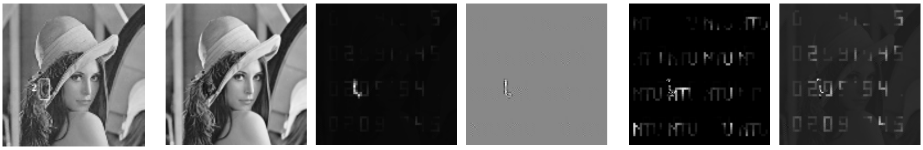
Fig. 5. Results of (a) A Gaussian noise test and (b) a multiplicative noise test.

3.2 Noise Addition Test

In the first test, different Gaussian white noises with the variance σ^2 ranging from 0 to 0.005 have been added to the watermarked image I^+ before undergoing the extraction. In the second test, a uniformly distributed random noise u with zero mean and variance σ^2 was multiplied and then added to the image by the equation $I^* = I^+ + uI^+$. The estimates watermarks were then compared to the originals and the results are shown in Fig. 5. The estimates extracted from the noisy image represent a high correlation with the original watermarks.

3.3 Authentication Test Results

From the watermarked image, a small area was copied and imperceptibly overwritten to another location (inside the white-bordered rectangle in Fig. 6(a)). The tampered image was then put into the extraction and we got three output images as shown in Fig. 6. The tampered area is clearly noticeable in both watermark estimates, the pixel values of the tampered area are much higher than the others. Moreover, if we replace those tampered pixels with the image average values, we can recognize the watermark. After the replacement and normalization, the estimated watermarks are clearly visible in Fig. 6(e) and 6(f).



(a) (b) (c) (d) (e) (f)

Fig. 6. (a) A down-sized version of the tampered image (the rectangle indicates the tampered area). (b)-(d) The three outputs: the down-sized estimates the original image $\hat{I}$, the first watermark $\hat{W}_1$ and the second watermark $\hat{W}_2$. (e)-(f) The two watermark estimates after the tampered area is removed: $\hat{W}_1^*$ and $\hat{W}_2^*$.

4 Conclusion

In this work, we have proposed a novel watermarking method which uses only one small-sized key image for extracting different watermarks. The two-watermark embedding technique allows us to verify the ownership of the image while keep tracking each copy of this work. Besides, the ICA-based extraction method makes possible to blindly estimate the watermarks without the original image.

Estimating the watermarks on size-reduced level is another advantage of the proposed algorithm. Even if the attackers, in some ways, might be able to extract the watermark's estimates, they are still unable to subtract these estimated watermarks from the image since the embedded watermarks and the watermark used to compare are of different sizes. Finally, simulations throughout various image attacks demonstrated the capability of the proposed method in both robust and fragile watermarking. The watermarks are difficult to remove and survive through the severe damage on the image. Moreover, malicious tampering on the image can also be detected during the extraction.

We are studying the ability of applying WMica into domain-transformed watermarking. The combination may result in a more robust watermarking technique. Further research to improve the performance may be carried on by employing different techniques for watermark generation, i.e., the function $\mathcal{M}(\cdot)$. Studying on visual masking $\mathcal{V}(\cdot)$ is also another interesting issue.

References

1. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital Watermarking. 1st edn. Morgan Kaufmann (2001)
2. Cichoki, A., Amari, S.: Adaptive blind signal and image processing. John Wiley & Sons Ltd (2002)
3. Zhang, S., Rajan, P.K.: Independent component analysis of digital image watermarking. In: Proc. of IEEE International Symposium on Circuits and Systems (ISCAS'02). Volume 3. (2002) 217–220
4. Gonzalez-Serrano, F.J., Molina-Bulla, H.Y., Murillo-Fuentes, J.J.: Independent component analysis applied to digital image watermarking. In: Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing, 2001 (ICASSP'01). Volume 3. (2001) 1997–2000
5. Shen, M., Zhang, X., Sun, L., Beadle, P.J., Chan, F.H.Y.: A method for digital image watermarking using ICA. In: Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Japan (2003) 209–214
6. Yu, D., Sattar, F., Ma, K.K.: Watermark detection and extraction using independent component analysis method. EURASIP Jounal on Applied Signal Processing – Special Issue on Nonlinear Signal and Image Processing **2002** (2002) 92–104
7. Voloshynovskiy, S., Herrigel, A., Baumgaertner, N., Pun, T.: A stochastic approach to content adaptive digital imagewatermarking. In: Proc. of International Workshop on Information Hiding, Dresden. Germany (1999) 212–236

Clustering-Based Image Retrieval Using Fast Exhaustive Multi-resolution Search Algorithm

Byung Cheol Song and Kang Wook Chun

Digital Media R & D Center, Samsung Electronics Co., Ltd,
416 Maetan-3dong, Yeongtong-gu, Suwon City, 443-742, Republic of Korea,
bcsong@samsung.com

Abstract. This paper presents a fast exhaustive multi-resolution search algorithm in a clustered image database. Prior to search process, the whole image data set is partitioned into a pre-defined number of clusters having similar feature contents. For a given query, the proposed algorithm first checks the lower bound of distances in each cluster, eliminating disqualified clusters. Next, it only examines the candidates in the surviving clusters through feature matching. Simulation results show that the proposed algorithm guarantees very rapid exhaustive search . . .

1 Introduction

Recently, content-based image retrieval (CBIR) has been strongly demanded for various applications. Low-level features such as colors, textures, and shapes are usually preferred as image contents [1-2].

In the conventional CBIR systems, if images are indexed by high-dimensional descriptors, the best match(es) for a query is searched by examining the similarity between the query descriptor and descriptors from database images, on high-dimensional space.

A straightforward way to perform similarity matching is the exhaustive search algorithm (ESA), where the full-resolution distances between the query and all candidate descriptors in the database are computed and compared each other. However, since the running time of ESA is proportional to the dimension (B) and size (N) of the dataset, i.e., $O(BN)$, ESA can be very costly. In order to logically speed up ESA, several fast exhaustive search algorithms using triangle inequality property or its variations have been proposed for various applications, e.g., motion estimation for video coding [3] and codebook search for vector quantization (VQ) encoding [4]. Berman and Shapiro employed the concept of triangle inequality to avoid full-resolution matching of unreliable candidate descriptors with the query descriptor, and reduce a great deal of computation [5]. However, its speed performance highly depends on the choice of key images, and is not satisfying in large image databases.

We propose a cluster-based search algorithm for fast exhaustive search. Prior to search process, entire images are grouped into a pre-defined number of clusters having similar contents through K-means clustering. Then, we derive a new

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 372–377, 2004.

search scheme based on two inequality properties. The first property is basically derived from the branch-and-bound search [6], which provides the lower bound of distances from the images in each cluster to a given query, and is used to eliminate disqualified clusters from the search procedure. The second one is from the distance relationship between adjacent levels on a multi-resolution feature space, alleviating unnecessary descriptor matching operations. The proposed algorithm using these two properties can determine the best match very rapidly.

The paper is organized as follows. In Section 2, we introduce two inequality properties and propose a cluster-based fast exhaustive search algorithm by systematically combining these two properties. Intensive experimental results are given in Section 3. Finally, conclusions are drawn in Section 4.

2 Cluster-Based Multi-resolution Search Algorithm

We classify an image data set into a pre-defined number of groups having similar color content by using the MacQueen K-means clustering algorithm. Let the k-th cluster and its center be Φ_k and C_k, respectively.

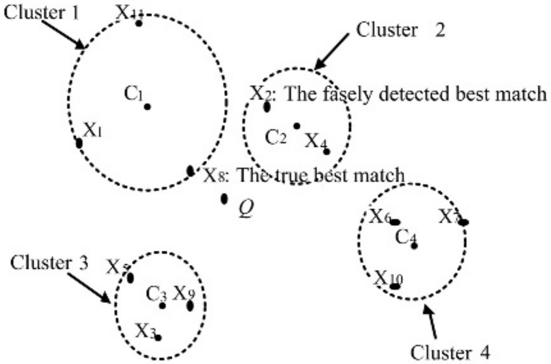


Fig. 1. Inherent problem of conventional cluster-based algorithms. Here, N=11 and K=4

2.1 Exhaustive Search Based on Branch-and-Bound Method

Fig. 1 demonstrates why conventional cluster-based algorithms could not achieve fast exhaustive search inherently. In the figure, X_i denotes the i-th data. Since C_2 is closest to the query Q among the cluster centers, the distances of all the elements in Φ_2 are examined, thereby X_2 is chosen as the best match. However, the true best match is X_8 in Φ_1. This situation arises because the true best match does not always exist in the cluster whose center is closest to Q. Although the best match is tried to retrieve from several clusters close to Q, this problem could not yet be solved thoroughly. In order to solve the above problem, the branch-and-bound search has been proposed for data search [6]. The main property in the branch-and-bound search is as follows:

$$\text{If} \quad d(C_k, Q) - \delta_k > d_{min}, \min_{X_i \in \Phi_k} d(X_i, Q) > d_{min} \tag{1}$$

where $d(X, Y)$ denotes the distance between two descriptors X and Y, d_{min} is "so far" minimum, and $\delta_k = \max_{X_i \in \Phi_k} d(X_i, C_k)$. Note that all δ_k's are pre-computed and stored. (1) can be proved easily from triangle inequality. By using d_{min} and all δ_k's, we can decide whether each cluster deserves to be examined or not for fast exhaustive search. In (1), $d(C_k, Q) - \delta_k$ means the lower bound of distances between Q and all the elements in the k-th cluster. Therefore, if $d(C_k, Q) - \delta_k$ is greater than d_{min}, this k-th cluster doesn't have to be considered anymore because it has no candidate whose distance is smaller than d_{min}.

2.2 Fast Exhaustive Search in Multi-resolution Database

To achieve fast exhaustive search for efficient image retrieval, we adopt the successive elimination algorithm (SEA) [3] to calculate the successive score on multi-resolution descriptors rather than multi-resolution images. Suppose that the dimension of an image descriptor is $B(= 2^L)$ and a multi-resolution structure for each descriptor X is defined as a descriptor sequence $X^0, X^1, \cdots, X^l, \cdots, X^L$, where $X = X^L$. The relation between X^i and X^{i+1} is as follows:

$$X^l(m) = X^{l+1}(2m - 1) + X^{l+1}(2m) \text{ for } 1 \le m \le 2^l, \tag{2}$$

where $X^l(m)$ denotes the m-th element value of X^l. Based on this multi-resolution descriptor structure, the following inequality property can be derived:

$$d(X, Y) \equiv d^L(X, Y) \ge \cdots \ge d^l(X, Y) \ge \cdots \ge d^0(X, Y), \tag{3}$$

where $d^l(X, Y)$ denotes the L_1-norm distance between two descriptors X and Y at level l, i.e., $d(X^l, Y^l)$. Note that L_1-norm distance is the most popular distance measure for image retrieval. (3) can be also proved easily. (3) indicates that if $d^l(X, Y)$ is larger than a certain value, the distances at upper levels are always larger than that particular value. So, if we apply this property to the search procedure, we can save a great deal of processing time by eliminating improper candidates at lower levels since the distance evaluation at upper levels needs much more computation time than it does at lower levels.

By taking advantage of property (3), we can achieve a fast exhaustive search. Assume that N denotes the total number of candidate images in the database and $X_1, \cdots, X_i, \cdots, X_N$ denotes the set of the corresponding descriptors. Multi-resolution descriptors of all images are pre-computed and stored. The image producing the final d_{min} is selected as the best match of a given query whose descriptor is Q. We refer to this fast exhaustive multi-resolution search algorithm as MSA.

2.3 Cluster-Based Multi-resolution Search Algorithm (CMSA)

By systematically combining the cluster-pruning scheme in subsection 2.1 and MSA in subsection 2.2, we propose a novel cluster-based multi-resolution search algorithm (CMSA) for image retrieval, which guarantees the same accuracy as ESA. Firstly, MSA is applied to find the nearest cluster center and the corresponding nearest candidate. Next, by using this information and property (1), cluster pruning is performed. Finally, MSA is applied again to search the best match in the survived clusters.

Since MSA is used to find the nearest cluster center, all $d_L(C_k.Q)$'s are not available in the following cluster pruning stage. So, we should modify property 1 by using the relation that $d(C_k, Q) \equiv d_L(C_k, Q) \geq d^{l_k}(C_k, Q)$, that is,

$$\text{If } \ d^{l_k}(C_k, Q) - \delta_k > d_{min}, \min_{X_i \in \Phi_k} d(X_i, Q) > d_{min} \tag{4}$$

In (4), l_k denotes the highest level where the computed distances are available and $l_k \leq L$. Since $d^{l_k}(C_k, Q)$'s and δ_k's are already known for all k, additional computation for the above decision is not required.

3 Experimental Results

We use a database containing 10,000 still images, i.e., N = 10,000: 7,000 images from an MPEG-7 content set [9] and 3,000 images from an ftp site. The database includes various types of images such as natural scenes, architectures, and people so as to prevent a bias to a particular type of image. The database is grouped into K clusters according to luminance histogram, and three cases of K=500, 1000, and 1500 are tried out. Besides database images, we use very different 100 still images as test images.

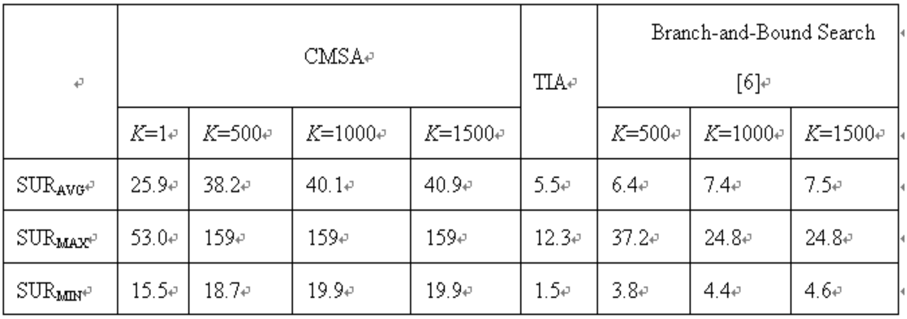
We also use a normalized luminance histogram as an image descriptor whose bin size B is 256. As an evaluation measure, we employ the speed-up ratio (SUR) in terms of CPU running time, which is defined as follows:

$$\text{SUR} = \frac{T_{ESA}}{T_{COMP}} \tag{5}$$

where T_{ESA} and T_{COMP} are the running time of ESA and the algorithm to be compared with, respectively. For fair comparison, we examine the following three kinds of SUR for 100 test images: average SUR (SUR_{AVG}), maximum SUR (SUR_{MAX}), and minimum SUR (SUR_{MAX}).

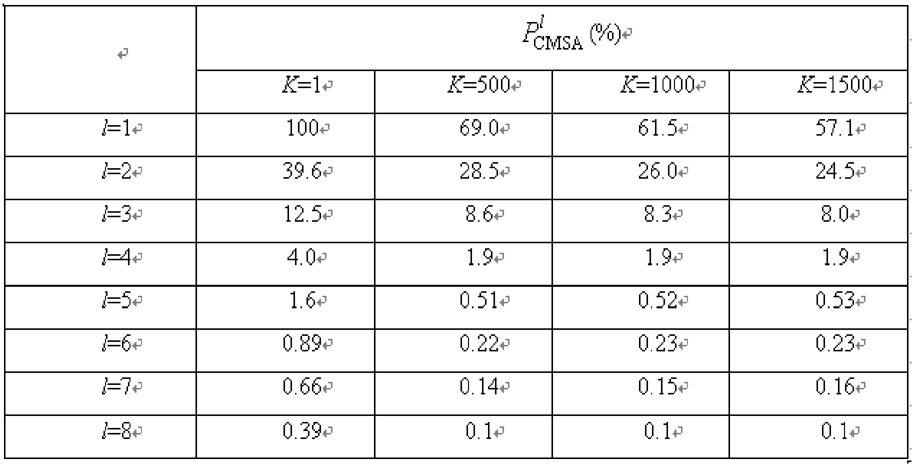
CMSA is compared with two existing algorithms: the branch-and-bound search algorithm [6] and a triangle inequality-based algorithm (TIA) [5]. Note that all the three algorithms guarantee the same search accuracy as ESA. In implementing TIA, we select 30 key images in random. The branch-and-bound search algorithm is implemented by employing the same K-means clustering as CMSA, instead of hierarchical clustering. In addition, an extreme case of K of 1 in CMSA is examined to show the effect of database clustering. Table 1 shows

Table 1. SUR comparison of CMSA for various M's and K's



	CMSA				TIA	Branch-and-Bound Search [6]		
	K=1	K=500	K=1000	K=1500		K=500	K=1000	K=1500
SUR_{AVG}	25.9	38.2	40.1	40.9	5.5	6.4	7.4	7.5
SUR_{MAX}	53.0	159	159	159	12.3	37.2	24.8	24.8
SUR_{MIN}	15.5	18.7	19.9	19.9	1.5	3.8	4.4	4.6

Table 2. Percentages of the candidates that are examined at each level of CMSA



	P^l_{CMSA} (%)			
	K=1	K=500	K=1000	K=1500
l=1	100	69.0	61.5	57.1
l=2	39.6	28.5	26.0	24.5
l=3	12.5	8.6	8.3	8.0
l=4	4.0	1.9	1.9	1.9
l=5	1.6	0.51	0.52	0.53
l=6	0.89	0.22	0.23	0.23
l=7	0.66	0.14	0.15	0.16
l=8	0.39	0.1	0.1	0.1

the comparison results for various K's. Let's check the case of a single best match ahead. When K is set in the range of 500 to 1500, CMSA is about 40 times faster than ESA on average, and is about 7.5 times faster than TIA on average, and also about 5.4 times faster than the branch-and-bound search. This is because the proposed algorithm employs an effective hybrid structure based on property (1) and property (3). On the other hand, note that CMSA speeds down by about 65It is also noticed that SURMIN of CMSA is always higher than the SUR_{MAX} of TIA.

Table 2 shows how many candidates are examined at each level of CMSA. The percentage of the candidates examined at each level is employed as an evaluation measure here. The percentage at level l is defined as follows:

$$p^l{}_{CMSA} = \begin{cases} \frac{\text{The number of candidates examined at level } l}{K+N}, \text{if } K \neq 1. \\ \frac{\text{The number of candidates examined at level } l}{N}, \text{otherwise} \end{cases} \quad (6)$$

In the clustered database, since CMSA deals with K cluster centers additionally during the search procedure, the denominator of ${p^l}_{CMSA}$ is to be $K + N$. Table 2 shows that CMSA statistically examines the candidates of 0.1 percent at the finest level when K=1500. In other words, the remaining candidates of 99.9 percent are eliminated at the coarser levels. Note that at coarser levels, ${p^l}_{CMSA}$ is smaller in the clustered database than in a database without clustering. This is the reason why CMSA provides high search speed in a clustered database.

4 Conclusions

This paper presented a fast exhaustive multi-resolution search algorithm. In the case of producing a single best match, the proposed algorithm is about 40 times faster than ESA, and is significantly faster than the existing algorithms. Even when producing multiple best matches, the proposed algorithm still provides remarkable speed performance. Although this paper only dealt with luminance histogram as an image descriptor, other descriptors with multiple dimensions can be also applied generally. Thus, the proposed algorithm is comprehensive and prospective for fast exhaustive search in large multimedia databases.

References

1. Pei, S.-C. and Cheng, C. -M.: Extracting color features and dynamic matching for image data-base retrieval, *IEEE Trans. Circ. and Syst. for Video Technol.*, **9** no. 3 (1999) 501–512
2. Manjunath, B. and Ma, W.: Texture features for browsing and retrieval of image data, *IEEE Trans. Pattern Anal. Machine Intell.*, **18** no. 8, (1996) 837–842
3. Li, W. and Salari, E.: Successive elimination algorithm for motion estimation, *IEEE Trans. Image Processing* **4** no. 1 (1995) 105–107
4. Gray, R. M. and Neuhoff, D. L.: Quantization, *IEEE Trans. on Information Theory* **44** no. 6 (1998) 2325–2383
5. Berman, A. P. and Shapiro, L. G.: Efficient image retrieval with multiple distance measures, *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases* **3022** (1997) 12–21
6. Fukunaga, K. and Narendra, P. M.: A branch and bound algorithm for computing K-nearest neighbors, *IEEE Trans. Computers*, (1975) 750–753
7. ISO/IEC JTC1/SC29/WG11/N2466: Licensing agreement for the MPEG-7 content set, Atlantic City, USA, (1998)

Robust Watermarking for Copyright Protection of 3D Polygonal Model

Wan-Hyun Cho[1], Myung-Eun Lee[2], Hyun Lim[3], and Soon-Young Park[2]

[1] Department of Statistics, Chonnam National University, South Korea
whcho@chonnam.ac.kr
[2] Department of Electronics Engineering, Mokpo National University, South Korea
{melee,sypark}@mokpo.ac.kr
[3] DVMM Lab., Dept. of Electrical Engineering, Columbia University, USA
hlim@ee.columbia.edu

Abstract. We describe a robust watermarking technique for proving ownership claims on 3D polygonal models. To make the proposed technique robust against a variety of attacks, a wavelet-based multiresolution analysis is used for a polygonal mesh model. First, we generate the simple mesh model and wavelet coefficient vectors by applying a multiresolution analysis to a given mesh model. Then watermark embedding is processed by perturbing the vertex of a chosen triangle mesh at a low resolution according to the order of norms of wavelet coefficient vectors using a look-up table. The watermark extraction procedure is to take the binary digits from the embedded triangle mesh using a look-up table and the similarity test between the embedded watermark and the extracted one is followed. The experimental results show that the proposed method is resistant to affine transformation such as scaling, translation and rotation of a 3D mesh model, as well as noise attacks.

1 Introduction

Digital watermarking is a technique designed to hide information in the original data for copyright protection or authentication of digital contents. Most of the researches have been focused on the watermarking technique for digital text, image, video, and sound signals up to now. Relatively, watermarking schemes for 3D models have been less concerned by researchers, since the existing techniques are not easily adapted to the arbitrary surfaces. But recently, several researchers have begun to have an interest about watermarking of 3D model as the 3D graphical objects or models are becoming much important in many areas of activity including digital cinematographys, virtual realitys as well as in CAD designs.

In general, watermarking techniques for multimedia data can be grouped into two categories; spatial domain methods and frequency domain methods. In the spatial domain watermarking for 3D polygonal model, the pioneer work has been conducted by Ohbuchi et al.[1]. They proposed several watermarking algorithms for polygonal models including Triangle similarity quadratic (TSQ),

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 378–385, 2004.

Tetrahedral volume ratio (TVR) and a visible mesh-watermarking algorithm. Benedens[2] subtly altered surface normal to embed watermark bits robust to the innocuous attacks. Yeo and Yeung[3] used a hash function to generate the cryptographic watermark signal and embed the watermark by perturbing the coordinates of vertices forming a given polygonal model. On the other hand, in the frequency domain watermarking method, Praun et al.[4] proposed the most successful robust mesh-watermarking algorithm that generalized the spread spectrum techniques to surfaces. Yin et al.[5] adopted Guskov's multiresolution signal processing method for meshes and used his 3D non-uniform relaxation operator to construct a Burt-Adelson pyramid for the mesh, and then watermark information was embedded into a suitable coarser mesh. Kanai et al.[6] used the wavelet transformation to obtain the multiresolution decomposition of polygonal mesh, and then embedded the watermark into the large wavelet coefficient vectors at various resolution levels of the multiresolution representation. Finally, Ohbuchi et al.[7] employed the mesh spectral analysis to modify mesh shapes in the transformed domain, and also Kohei et al.[8] computed the spectrum of the vertex series using the singular spectrum analysis for the trajectory matrix derived from the vertices of 3D polygonal mesh, and then embedded the watermark into the singular values given from spectrum analysis.

In this paper, we propose a new robust and blind watermarking algorithm for given 3D polygonal models using the wavelet-based multiresolution analysis. We employ the multiresolution analysis for a given polygonal model using the subdivision method proposed by Eck et al.[9] and derive a simple mesh that is topologically equivalent to the given polygonal mesh model as well as wavelet coefficient vectors. Watermark embedding is processed by perturbing the vertex of a chosen triangle mesh at a low resolution according to the order of norms of wavelet coefficient vectors using a look-up table(LUT). On the other hand, the watermark extraction procedure is to take the binary digits from the embedded triangle mesh using an LUT and the similarity test between the embedded watermark and the extracted one is followed.

The structure of this paper is organized as follows. In Sect. 2, we describe the watermark embedding and extracting algorithm. Sect. 3 discusses the properties of detector statistic and the experimental results are given in Sect. 4. Finally, the conclusions are mentioned in Sect. 5.

2 Watermarking Algorithms for 3D Model

We propose a new watermarking technique to protect illegal copy of 3D polygonal mesh models. The following block diagram shows the general structure for our watermarking scheme. On the watermark embedding stage, our scheme first conduct the multiresolution analysis for the original polygonal mesh by using lazy wavelet transform and then determine a simple mesh and wavelet coefficient vectors. Next, we insert the watermark into the coordinate of a vertex using an LUT. After that, we apply the inverse wavelet transform to the modified simple mesh and the unchanged wavelet coefficients vectors to construct the watermarked polygonal mesh. On the other hand, the order of extracting the watermark is processed as a reverse manner with the embedding procedure and the correlation test between the embedded watermark and extracted one is followed to prove the ownership claims on the original polygonal mesh.

2.1 Watermark Embedding Procedure

Multiresolution analysis for 3D polygonal model. By applying wavelet transform to original polygonal mesh, we derive both a simple mesh model and wavelet coefficient vectors. If we conduct the multiresolution analysis using the subdivision method for arbitrary 3D polygonal meshes at several times, we can ultimately obtain a coarse mesh that is topologically equivalent to the given polygonal mesh and a collection of wavelet coefficients. Here this model is called a simple mesh or a base mesh. Fig. 1(a) shows the Venus model and Fig. 1(b) shows the base mesh that was produced by applying Eck[9]'s algorithm for this model.

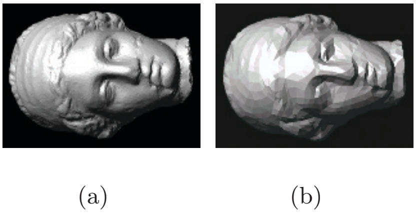

(a) (b)

Fig. 1. Multiresolution analysis about Venus model, (a) Venus model, (b) simple mesh constructed by multiresolution analysis

Embedding step of the watermark into a coordinate of a mesh vertex. We decompose the 3D polygon by using lazy wavelet transform and choose an arbitrary triangle, t_i^j at multiresolution level j with three wavelet coefficient vectors, w_{i1}^j, w_{i2}^j, w_{i3}^j and three vertices a, b, c as shown in Fig. 2.

Next, we compute the sum of norms of three wavelet coefficient vectors and sort all these sums with their magnitudes to insert the watermark into the vertex of a triangle. Triangles are selected as the same number as watermark bits by ordering the sums of norms from a simple mesh model and decide a vertex in the selected triangle to embed a watermark. The fundamental idea in this embedding procedure is to modify the vertex in the rugged region with large wavelet coefficient vectors in order to avoid the perceptual visibility.

To embed the watermark, we need to construct an LUT using the information extracted from the simple triangle mesh. First, we select an edge corresponding to the wavelet coefficient vector having the largest norm. Then we draw the straight line from the middle point of the selected edge to the opposite vertex and compute the ratio of the length of the new straight line to the length of the selected edge.

$$Ratio = \frac{||c - \frac{1}{2}(a+b)||}{||a-b||} \tag{1}$$

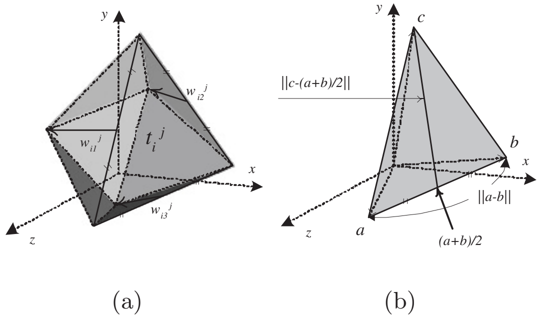


Fig. 2. A typical triangle selected from a simple mesh, (a) three wavelet coefficient vectors, (b) geometrical diagram for watermark embedding

Next, we partition a distributed interval of all ratios computed from given triangles into K subintervals (bin), and generate a sequence of binary random digits composing with "0" and "1" seeded with a secret key. The LUT is constructed as we allocate each digit to each bin one by one. After constructing the LUT, we compute a ratio from a selected triangle and take a binary digit corresponding to this ratio using the LUT. In addition, we take a binary digit of a watermark using the computed location index. Finally, we compare two binary digits and then if two digits are equal, we don't have any change about the vertex coordinate. But if two digits are not equal, we select the nearest neighboring subinterval so that the digit of the LUT equals to the watermark digit. The new ratio coming from the selected subinterval is then used to move the vertex coordinate c up and down.

2.2 Watermark Extraction Procedure

To extract an embedded watermark from a 3D polygonal mesh, we first apply the wavelet transform to the observed mesh and then generate the watermarked simple mesh and a collection of wavelet coefficient vectors. Second, we choose an arbitrary triangle from simple mesh model, and then compute the sum of norms of three wavelet coefficient vectors related with the triangle. After ordering the sums of norms according to their magnitudes we select the triangle with the embedded vertex and compute a ratio of the height to the length of the bottom edge of a selected triangle. Now we can restore the watermark by picking up a binary digit from the LUT corresponding to the computed ratio. Finally, we compute the correlation statistic between the original watermark and the restored watermark to verify the illegal copy of 3D polygon mesh.

3 Watermark Detection Using Statistical Measure

For the copyright protection on 3D polygonal model the statistical approaches are widespread. Here we will use a statistical similarity measure between the extracted watermark and the embedded watermark. First, we define the Bernoulli variables using two kinds of watermarks as the following manner. Suppose that random variable, W_i^* is the watermark signal extracted from watermarked polygonal model and random variable, W_i is the embedded watermark. If two random variables are equal, then the variable X_i takes a value, ”1.” Otherwise, this takes a value ”0.” Then this variable becomes the Bernoulli variable that the success probability is given as:

$$p = P(X_i = 1) = P(W_i = W_i^*). \tag{2}$$

Hence, if we use the N Bernoullii variables derived from each ratio, we can obtain the following detector statistics.

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{3}$$

Then, this statistic is a sample proportion of watermark detection probability. Thus, we can verify the ownership of 3D polygonal model using the sample statistic properties. In order to verify this problem statistically, we first define the two kinds of hypotheses.

H_0: The host media does not contain the claimed watermark: $W_i^* = N_i$

H_1: The host media contains the claimed watermark:$W_i^* = W_i + N_i$
where N_i is binary noise.

Here, if the null hypothesis is true, then the expectation and variance of detector statistic are given by

$$\begin{aligned} \mu_{H_0} &= E(\hat{p}) = E[\tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i] = p_0 \\ \sigma_{H_0}^2 &= V(\hat{p}) = V[\tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i] = \tfrac{p_0(1-p_0)}{N}. \end{aligned} \tag{4}$$

Then, under the null hypothesis, the success probability is given by

$$\begin{aligned} p_0 &= P(W_i = W_i^*) = P(W_i = N_i) \\ &= P(W_i = 0, N_i = 0) + P(W_i = 1, N_i = 1) \\ &= P(W_i = 0)P(N_i = 0) + P(W_i = 1)P(N_i = 1) = \tfrac{1}{2} \end{aligned} \tag{5}$$

Here, the third equality will be hold because of independency of two kinds of watermarks. On the other hand, if the alternative hypothesis is true, then the expectation and variance of detector statistic are given as different form.

$$\begin{aligned} \mu_{H_1} &= E(\hat{p}) = E[\tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i] = p_1 \\ \sigma_{H_1}^2 &= V(\hat{p}) = V[\tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i] = \tfrac{p_1(1-p_1)}{N}. \end{aligned} \tag{6}$$

Then, under the alternative hypothesis, the success probability is given by

$$\begin{aligned} p_1 &= P(W_i = W_i^*) = P(W_i = W_i + N_i) \\ &= P(W_i = 0, W_i + N_i = 0) + P(W_i = 1, W_i + N_i = 1) \\ &= P(W_i = 0)P(W_i + N_i = 0|W_i = 0) + P(W_i = 1)P(W_i + N_i = 1|W_i = 1) \\ &= \tfrac{1}{2} + f_n(\cdot) \end{aligned} \tag{7}$$

where $f_n(\cdot)$ is a function of a relative ratio of an original bin length to a reduced bin length due to the noise addition. For example, suppose a length of bin is equal to L and if a $\alpha\%$ noise is inserted into the triangle mesh, then a variation of bin length is $l = L \times 0.1\alpha$ and the probability that a watermark may be changed is given by $\frac{2l}{L}$. So, the probability which doesn't change is $1 - (\frac{2l}{L})$ and $f_n(\cdot)$ is proportional to $\frac{(L-2l)}{2L}$. And also, if we use the central limit theorem for a sample proportion, the distribution of detector statistic is approximated with the normal distribution under null hypothesis as well as the alternative hypothesis. So, the detecting probability of the true watermark which is rejection of null hypothesis is given by

$$DP = P(\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} \geq c). \tag{8}$$

And the missing probability which doesn't detect the true watermark is given as

$$MP = P(\frac{\hat{p} - p_1}{\sqrt{\frac{p_1(1-p_1)}{N}}} \leq c). \tag{9}$$

Hence, we have to find the critical value, c which both can make the detecting probability as big as possible but also the missing probability as small as possible.

4 Experimental Results

We have applied our algorithm to the 3D Venus model as shown in Fig. 1(a). This model consists of 33,537 vertices and 67,072 faces. Then this model is decomposed into both a simple mesh and wavelet coefficient vectors for a multiresolution analysis by using the lazy wavelet 4–to–1 subdivision scheme. That is, the polygon mesh has the mesh topology which fits to 4–to–1 subdivision scheme. The two level decomposition yields the simple mesh model with approximately 2,097 vertices. After ordering the sums of the norms of wavelet coefficients vectors, 200 binary digits generated by random sequence are embedded into 2,098 vertices with 4,192 faces. Fig. 3(a) shows the reconstructed watermarked polygon mesh after embedding watermarks to the selected 200 faces of the simple mesh. It can be observed that the watermarked polygon mesh is imperceptible after embedding the watermark into the vertices in the rugged region. Fig. 3(b)–(d) show the watermarked mesh after applying Gaussian distributed noise of variance $\nu = 0.3\%$ per coordinate, after applying 50% scaling, and after applying rotation operation.

We have tested the embedded key and 99 random keys for the watermarked Venus model with the normalizing correlation results as Fig. 4. The detector yielded the maximum correlation value for the correct key and the correlation values decreased for the use of wrong keys. We also investigated the effect of the noise attack and the affine transformation on the watermarked model, and the correlation result of the false alarm condition. Fig. 5 shows the distribution of correlation for 500 tests. It is obvious that the proposed technique has shown to be robust to the affine transformation and noise attacks by selecting the proper threshold value.

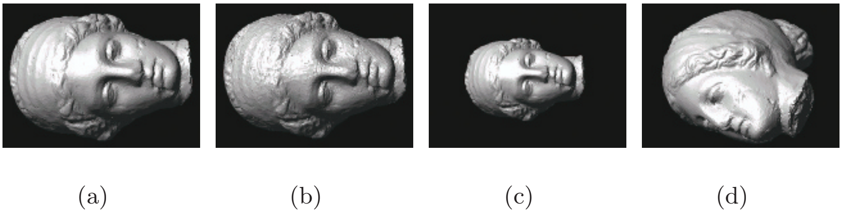


Fig. 3. (a) The watermarked Venus model, (b) the noisy watermarked model, (c) 50% scaled watermarked model, and (d) rotated watermarked model

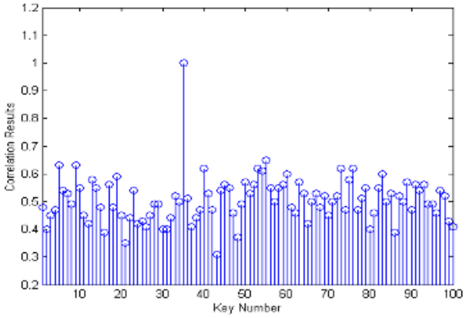


Fig. 4. Test results for 100 random keys

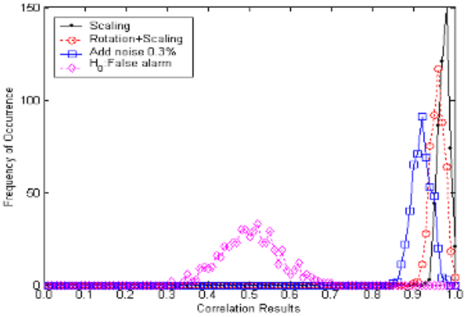


Fig. 5. Distribution of the probability of false alarm and the detecting probability for an attacked model

5 Conclusion

In this paper, we proposed the robust watermarking technique for 3D polygonal mesh model. Our method is based on the multiresolution analysis for arbitrary 3D polygon using wavelet transform. The proposed algorithm employs a wavelet-based multiresolution analysis to convert the original polygonal mesh model into a simplified mesh model and wavelet coefficient vectors. We select vertices from the simplified mesh model according to the order of sums of wavelet coefficient norms and insert watermarks into them using a look-up table. The perceptual invisibility of the proposed technique is provided by embedding the watermark into the rugged region with large wavelet coefficient vectors and the invariance to the affine transformation is provided by employing the invariant properties of both the norm of wavelet coefficient vectors and a ratio of two lengths between the selected vertices. The experimental results have showed that the proposed technique can be a powerful copyright protection method of 3D polygonal meshes by surviving to the innocuous attacks and noise attacks.

References

1. Ohbuchi, R., Masuda, H., Aono, M.: Watermarking Three Dimensional Polygonal Models. Proceedings of ACM Multimedia, Seattle, USA (1997) 261–272
2. Benedens, O.: Watermarking of 3D polygonal based models with robustness against mesh simplication. Proceedings of SPIE: Security and Watermarking of Multimedia Contents, San Jose, USA (1999) 329–340
3. Yeo, B. L., Yeung, M. M.: Watermarking 3-D Objects for Verification. IEEE Computer Graphics and Application, Vol. 19, (1999) 36–45
4. Praun, E., Hoppe, H., Frankelstein, A.: Robust Mesh Watermarking. Proceedings of SIGGRAPH'99, LA, USA (1999) 49–56
5. Yin, K., Pan, Z., Shi, J., Zhang, D.: Robust Mesh Watermarking Based on Multiresolution Processing. Computer and Graphics, Vol. 25, (2002) 409–420
6. Kanai, S., Date, H., Kishinami, T.: Digital Watermarking for 3D polygons using Multiresolution Wavelet Decomposition. Proceedings of Sixth IFIP WG 5.2 GEO-6, (1998) 296–307
7. Ohbuchi, R., Mukaiyama, A., Takahashi, S.: A Frequency Domain approach to Watermarking 3D Shapes. Computer Graphics Forum, Vol. 21, (2002) 373–382
8. Kohei, M., Kokichi, S.: Watermarking 3D Polygonal Meshes Using the Singular Spectrum Analysis. Mathematics of Surfaces: 10th IMA International Conference, Leeds, UK (2003) 85–98
9. Eck, M., DeRose, T., Duchmap, T., Hoppe, H., Lounsbery, M., Stuetzle, W.: Multiresolution analysis of arbitrary meshes. In Proceedings of SIGGRAPH'95, ACM New York (1995) 173–182

Digital Video Scrambling Method Using Intra Prediction Mode

Jinhaeng Ahn[1], Hiuk Jae Shim[2], Byeungwoo Jeon[3], and Inchoon Choi[4]

School of Information and Communication Engineering, Sungkyunkwan University
300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, Republic of Korea.
{[1]skajh,[2]waitnual,[4]sonne}@ece.skku.ac.kr,[3]bjeon@yurim.skku.ac.kr

Abstract. As the amount of digitalized contents increases rapidly, 'security' necessarily arises as one of the most important issues. The main distribution channel of digital contents is Internet which is very easily accessible. Therefore content protection becomes a major issue as important as data coding techniques. In recent years, many developers have studied techniques that allow only authorized person to access contents. Among them, scrambling is one of well-known security techniques. In this paper, we propose a simple and effective digital video scrambling method which utilizes the intra block properties of a recent video coding technique, H.264. In addition to its simplicity, the proposed method does not cause bit rate increase after scrambling.

1 Introduction

As digital contents have been widely produced, security of multimedia data is highly required. By its nature, the feature that one can easily copy and distribute digital content has been believed to be advantageous for a while; however, it is no longer just "advantageous" in these days. Especially digital video sequences are widely distributed through non-private channel such as satellite links, cable television networks, wireless networks, and the Internet. Accordingly, video content providers demand more secure but simple techniques such as video scrambling. In general, video scrambling is a well-known analog technique that protects cable television signal from unauthorized user. In this paper, the main concept of scrambling is translated into the digital form of video signals, and digital video scrambling techniques applicable to MPEG-1, 2, 4 have been proposed. A digital video scrambling method distorts video signal as much as one can barely recognize original one. Only an authorized receiver who has 'key' or a descrambler can properly restore the original video signal. There is an additional purpose that an unauthorized receiver decodes incorrectly a scrambled video sequence. Despite distorted visual quality, one can still recognize important objects such as actor's movement which stimulates an unauthorized viewer's interest. In recent years, many researchers have proposed various different scrambling techniques [1]- [5].

The motion vector scrambling method utilizes CBP (Coded Block Pattern) information and motion vectors [5]. In inter frame coding, VLC code of the given

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 386–393, 2004.

motion vector is changed in accordance to the value of CBP modula 33. However, it has main disadvantage of increased bit rate of scrambled video.

Another existing scrambling method utilizes frequency domain such as wavelet or DCT [1]. There are many wavelet transform-based scrambling techniques such as selective bit scrambling, block shuffling, block rotation, etc., however, most of video coding methods such as MPEG-1, 2, 4, and H.264 do not adopt wavelet transform process. As a result, an additional wavelet transform process is required to scramble and descramble video sequences, and these methods have the same drawback of increased bit rate after scrambling. There are also many scrambling techniques based on 8x8 DCT transform. Those are DCT coefficient scrambling, motion vector scrambling, sign encryption, and so on. Most of these techniques also increase the bit rate of video sequence like other existing techniques [2]. Among them, the sign encryption [1] is one which does not increase bit rate, however, computational overheads are considerably increased. Since it changes the sign of every DCT coefficient, coding complexity is increased by up to about 15-20% [1]. Therefore we can conclude that both bit rate increase and complexity are the major problems of video scrambling techniques. Under the constraints, we propose a new scrambling method utilizing an intra block coding scheme. The proposed method satisfies its original purpose that it distorts visual quality but still provides minimal information which can induce interest of unauthorized receiver at the same time. However the proposed technique does not increase bit rate, and it is efficient and easy to implement. Also, there is little increase of complexity increases. Therefore it will be an attractive method on providing digital video.

The proposed method is described in Section 2, and its experimental results are shown in Section 3. Then we summarize the whole procedure and draw conclusion in Section 4.

2 Proposed Scrambling Method

The H.264 video coding standard employs intra prediction technique in order to remove spatial redundancies within intra frame. To decide the best intra prediction, it calculates SAD (Sum of Absolute Difference) as prediction error along specified directions, and a mode that has the smallest error is decided as the chosen intra prediction mode. After calculating the intra prediction mode, prediction residual values are obtained by subtracting the pixel values of selected prediction mode from current block. Then both the residual values and the prediction mode are transmitted to entropy encoder. Since video sequences can be easily distorted by modification of the prediction mode, the intra prediction mode is our main target for easy video scrambling. Accordingly the proposed scrambling method is based on intra block coding. In addition to the simple modification of intra block, another reason of considering only intra block is the effect of error propagation. The first frame of every video sequence is encoded with the intra coding technique and following inter frames refer the intra frame during inter coding. Therefore if we scramble only intra frame, inter frames undergo distortion propagated from the scrambled intra frame. Due to this error propagation, we don't have to scramble every inter frame.

The Intra coding scheme in H.264 has two cases: The Intra 4x4 and the Intra 16x16.

In case of the Intra 4x4, the size of unit block is 4x4 and there are 9 directions of intra prediction mode. Since the 4x4 prediction modes consist of 9 cases, at least 4 bits are required to encode the selected mode correctly. However, only 3 bits are used by utilization of a flag bit. It is called 'prev_intra4x4_pred_mode'. If the prediction mode of current block is equal to minimum prediction mode of its two neighboring upper and left blocks, the flag bit is set to '1'. Otherwise, the flag bit is set to '0'. By signaling the flag bit, we can exclude one mode which is equal to the minimum value from 9 modes. Therefore only 8 modes are to be encoded, which means only 3 bits are enough to represent every mode. Thus Intra 4x4 prediction mode is encoded with 3 bits fixed length code.

In the Intra 4x4 prediction, every prediction mode is calculated with neighboring upper or left block of current block. For instance, if only upper block is available, then 'horizontal', 'diagonal down/right', 'vertical right', 'diagonal down/left', and 'horizontal up' modes can be chosen. Therefore, in case that either upper or left block is not available to current block, the number of available mode can be reduced. However, the proposed scramble method modifies prediction modes only when both neighboring blocks are available. Since both blocks are available in most cases, it is enough to scramble video sequence, and the consideration of neighboring block increases its complexity. However, if more visual distortion is required, it is also possible to consider every case of existing block such as 'both available', 'only upper available', 'only left available', and 'both unavailable'.

When the flag bit is '1', the encoder does not send the prediction mode and no scrambling is done in this case, but when the flag bit is '0', intra prediction modes are modified. For scrambling, pseudo random sequence is generated by a given specific key. If the flag bit is "1", 3 bits are read from the pseudo random sequence. And then we obtain new prediction mode with following equation.

$$\text{Mode}_{\text{new}} = \text{Mode}_{\text{org}} \oplus 3 \text{ bits random sequence}\,, \tag{1}$$

where $\oplus$ is XOR (exclusive-or) operator. Since the length of Mode_{new} and the original prediction mode are the same, there is no bit rate increase after the proposed scrambling procedure. In addition, the exclusive-or operator is very simple to implement, and complexity of the operator is trivial.

Similarly in case of Intra 16x16, the size of unit block is 16x16 and there are 4 intra prediction modes. However, Intra 16x16 prediction modes are encoded in a different way from the Intra 4x4 case. In H.264 standard, the 16x16 prediction modes are encoded as variable length code instead of fixed length code. Moreover the Intra 16x16 prediction modes are jointly coded with luma and chroma CBP (Coded Block Pattern) values. Accordingly, the VLC table is utilized to encode the Intra 16x16 prediction mode. Since luma and chroma CBP is jointly coded with the prediction mode, the Intra 16x16 prediction modes can not be modified in the same way as Intra 4x4 are. When they are to be modified, the CBP values should not be changed; otherwise video sequences can not be decoded properly. Also bit rate should be preserved after scrambling. A simple way to satisfy these conditions is to find a pair of possible exchanging modes. For example, mode 0 and mode 1 in the VLC table shown in Table 1 constitutes a pair. Mode 2 and mode 3 is the other pair, and so on. Thus in case of the Intra 16x16, only one bit is required from the pseudo random sequence. If the bit is 1, then the parity bit of prediction mode is changed. Otherwise, the original prediction mode is not changed. This means that the mode is modified to its pair mode only when pseudo random sequence is 1. As one can see in Table 1, even though we change a mode to its pair mode, CBP values remain unchanged. Beside, the code length of the new mode and the original mode is the same. For example, since the pair of mode 1 is mode 0, when the pseudo random bit is 1 and the original mode is 0, mode 0 is modified to mode 1. Whatever the macroblcok type is, it is apparent that CBP value and the code length of both new mode and original mode are not changed as can be confirmed in Table 1.

Table 1. VLC table for Intra 16x16 prediction mode

mb_type	Name of mb_type	Intra16x16 PredMode	CodedBlock PatternChroma	CodedBlock PatternLuma	Code Legth	Pair Number
1	I_16x16_0_0_0	0	0	0	3	0
2	I_16x16_1_0_0	1	0	0	3	0
3	I_16x16_2_0_0	2	0	0	5	1
4	I_16x16_3_0_0	3	0	0	5	1
5	I_16x16_0_1_0	0	1	0	5	2
6	I_16x16_1_1_0	1	1	0	5	2
7	I_16x16_2_1_0	2	1	0	7	3
8	I_16x16_3_1_0	3	1	0	7	3
9	I_16x16_0_2_0	0	2	0	7	4
10	I_16x16_1_2_0	1	2	0	7	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Since only the parity bit of a prediction mode is changed, there is no increase of complexity. And when the generated pseudo random sequence exhausted, we circulate the sequence to the first bit, which means that we read the required bits again from the beginning of the sequence.

De-scrambling procedure is similar to the scrambling procedures. In case of the Intra 4x4, 3 bits are read from the given pseudo random sequence. Authorized receivers are supposed to have the same ‘key’ used by the scrambling procedure, therefore they can generate the pseudo random sequence. After reading 3 bits from the pseudo random sequence, the original video can be reconstructed with exclusive-or operation. This is as follows.

$$\mathrm{Mode}_{\mathrm{org}} = \mathrm{Mode}_{\mathrm{new}} \oplus 3 \text{ bits random sequence} \tag{2}$$

The property of exclusive-or operator guarantees that original prediction mode is properly obtained.

In case of the Intra 16x16, 1 bit is read from the pseudo random sequence. If the bit is 1, then is reverted the parity bit of received mode. ‘The bit is 1’ means that transmitter modifies an original mode to its pair mode. Therefore, we convert a received mode to its pair again. When the read bit is 0, we use a received mode as prediction mode. Consequently, only the authorized receiver with proper seed can reconstruct a scrambled video sequence, which is the same as the case of Intra 4x4. Similarly there are no increase in the size of bitstream and complexity. Scrambling and de-scrambling pseudo-codes are as follows.

```
Scrambling Pseudo-code

   switch (Intra mode)
      case intra 4x4:
         if prev_intra4x4_pred_mode != 1
            read 3 bits from pseudo random sequence
            new_mode = original_mode XOR 3_bits
         else
            new_mode = original_mode
         transmit new mode
      case intra 16x16:
         read 1 bit from pseudo random sequence
         if the bit == 1
            change parity bit of original mode
         transmit the mode

De-scrambling Pseudo-code

   switch (Intra mode)
      case intra 4x4:
         if prev_intra4x4_pred_mode != 1
            read 3 bits from pseudo random sequence
            original_mode = received_mode XOR 3_bits
      case intra 16x16:
         read 1 bit from pseudo random sequence
         if the bit == 1
            change parity bit of received mode
```

3 Experimental Results

For the simulation of the proposed video scrambling, well-known sequences such as ‘paris’, ‘mother and daughter’, and ‘foreman’ sequences are used. The size of these sequences is CIF. Note that ‘Foreman’ sequence has camera panning, therefore, it hse only small number of intra blocks in inter frames. ‘Paris’ and ‘mother and daughter’ sequences contain complex and simple background, respectively. Since these backgrounds do not have motions, it remains as it is scrambled. All sequences are simulated with H.264 reference software JM (Joint Model) 8.1a under the baseline profile. Only the first frame of each sequence is encoded as Intra frame.

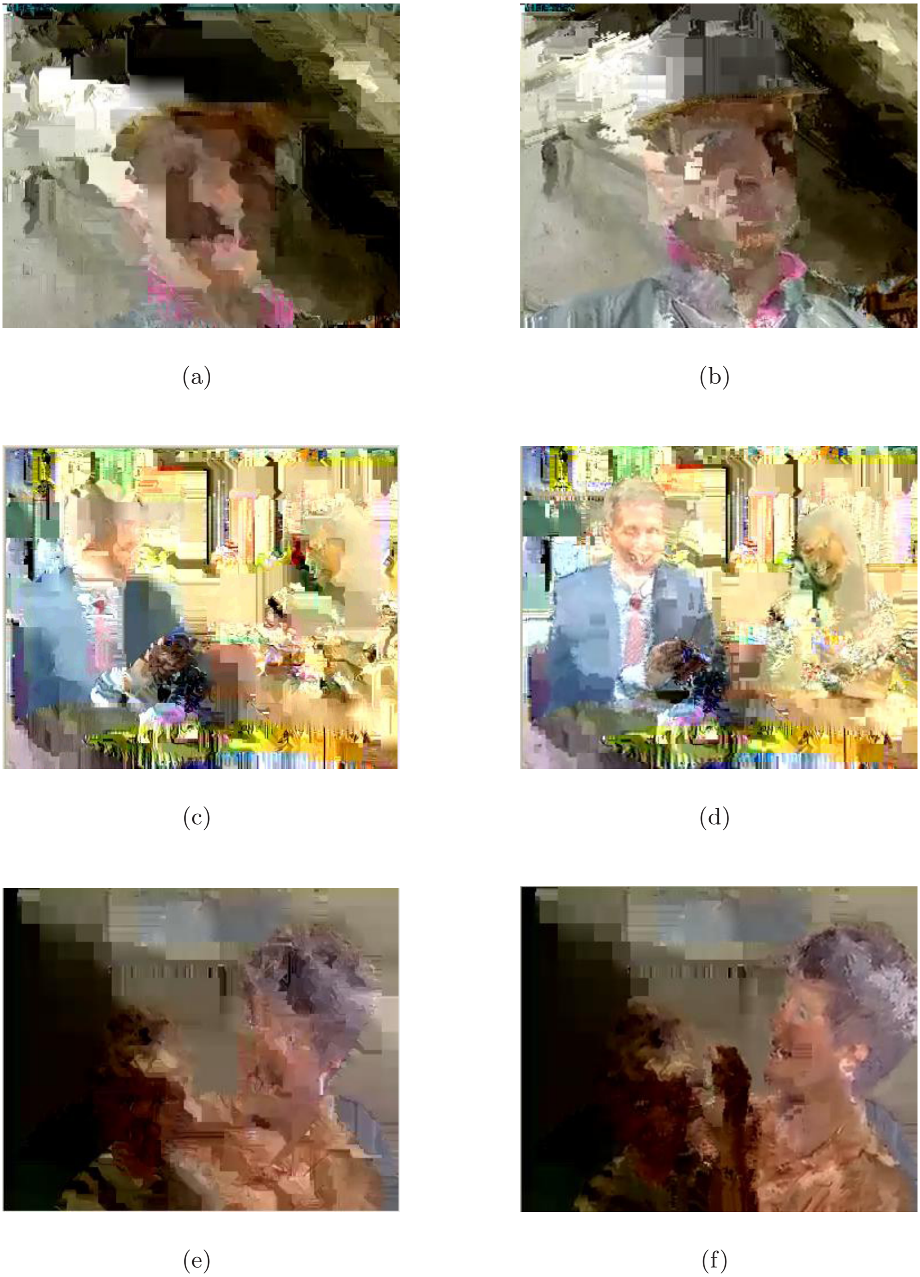
The scrambled sequences of the proposed method are shown in Fig. 1. For the comparison of intra and inter frames, the first frame and the 100th frame are shown. All blocks in the first frame is scrambled. All other fames are encoded as inter frame. Therefore, only small number of intra blocks in the inter frames are scrambled. Thus inter frames are affected only by error propagation from the scrambled intra blocks mostly at the first intra frame. The effect of error propagation works as similar as the scrambling effect. This can be observed from Fig. 1 (b), (d), and (f).

4 Conclusion

We proposed a scrambling method using intra prediction mode in H.264. The proposed method is designed to scramble only the intra blocks, therefore there is no direct scrambling in inter blocks. The second inter frame, however, refers to scrambled intra blocks in the first frame as referencem. Therefore the second frame can not be reconstructed correctly. In the same manner, following successive frames refer to distorted blocks in their previous frames. Thus, scrambling errors of intra blocks are propagated through video sequence. As a result, we can obtain scrambling effects as much as intra blocks without direct inter scrambling. Therefore we can reduce computational overheads considerably in contrast to other scrambling methods that modify both intra and inter blocks. In addition, bit rate is not increased at all.

The idea of our method is the simple modification of intra prediction mode without bit overheads, and is verified with H.264 video coding scheme. However this concept can be applied to other video coding schemes that adopt similar intra prediction techniques.

The purpose of scrambling is to allow only authorized receiver to access the original video sequence. The proposed method achieves the goal without bit overheads. For correct descrambling, a pre-defined key is delivered to the authorized receiver. With the proposed method, video sequence is not totally distorted, which makes unauthorized receivers are still capable of recognizing slight movements which stimulate the interest of unauthorized receivers from scrambled sequences.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. The first and 100th decoded frames using the proposed scrambling method: (a) the first frame of ‘foreman’ sequence, (b) 100th frame of ‘foreman’ sequence, (c) the first frame of ‘paris’ sequence, (d) 100th frame of ‘paris’ sequence, (e) the first frame of ‘mother and daughter’ sequence, (f) 100th frame of ‘mother and daughter’ sequence

References

1. W. Zeng and S. Lei: Efficient frequency domain selective scrambling of digital video IEEE Transactions on Multimedia, March (2003) 118–129
2. B. Macq and J. Quisquate: Digital images multiresolution encryption Interactive Multimedia Assoc. Intell. Property Proj., Jan. (1994) 179–186
3. W. Zeng and S. Lei: Efficient frequency domain video scrambling for content access control ACM Multimedia '99, Nov. (1999)
4. N. Katta et al. Scrambling apparatus and descramble apparatus U.S patent 5377266, Dec. 27 (1994)
5. J. Junpil: Digital video scrambling method KR patent 0151199, Jun. 18 (1998)

Watermark Re-synchronization Using Sinusoidal Signals in DT-CWT Domain

Miin-Luen Day[1,3], Suh-Yin Lee[1], and I-Chang Jou[2]

[1] Department of Computer Science and Information Engineering,
National Chiao-Tung University, Hsin-Chu, Taiwan, R.O.C.
sylee@csie.nctu.edu.tw

[2] Department of Computer and Communication Engineering,
National Kaohsiung First University of Science and Technology, Kaohsiung,
Taiwan, R.O.C.
icjou@ccms.nkfu.edu.tw

[3] Telecommunication Laboratory, Chunghwa Telecom Co., Chung-Li, Taiwan, R.O.C.
day@cht.com.tw

Abstract. Embedding sinusoidal signals or tiles patterns into image in the spatial domain to form some peaks is an effective technique for geometric invariant image watermark detection. However, there are two drawbacks in these spatial domain based schemes: one is poor picture quality of resulting watermarked image and the other is weak peaks visibility which is hard to detect. The previous works suffer from requiring a very strong watermark embedding to ease re-synchronization peaks finding, which in turn leads to a poorly watermarked image. To overcome this problem, we explore embedding sinusoidal signals individually in each of the selected sub-bands of dual-tree complex wavelet transform domain (DT-CWT), and then detecting the re-synchronization peaks by using the accumulated embedding sub-bands. Experimental results demonstrate that by adopting our approach, the resulting re-synchronization peaks are robust to rotation, scaling and translation attacks while preserving the visual quality of the watermarked image, thereby resolve the unavoidable dilemma faced by the other schemes.

1 Introduction

Watermarking is a technique to hide data or information imperceptibly within image, audio or video so that valuable contents can be protected. Though progress of watermark embedding / detection techniques has been continuously made through active devotion by many researchers, the attack counterpart increases in the mean time. There exists no sole watermark solution resistant to all types of attacks. Among many types of attacks, geometric attack remains one of the most challenging and unsolved problem, and more and more algorithms are devoted to combating geometric attacks. There are primarily two main categories of watermarking techniques for resolving geometric attacks in the literature: one approach is by using invariant features, and the other approach is by adopting re-synchronization template. For re-synchronization approach [1, 2], Kutter

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 394–401, 2004.

[1] embeds the self-reference pattern several times at horizontally and vertically shifted locations for recovering affine transformations. Fleet and Heeger [2] embeds sinusoidal signals into color image instead grey one to avoid artifact of watermarked image.

It is well recognized that the conventional discrete wavelet transform (DWT) is not suitable for geometric invariant watermark detection due to their lacking of shift-invariance property. Kingsbury [3] proposed the dual-tree complex wavelet transform (DT-CWT) which has both the advantages of being approximately shift-invariant and having additional directionalities (+15, +45, +75, -15, -45 and -75) compared to 3 directionalities (H, V, D) for traditional DWT. There are many successful applications by using DT-CWT, such as motion estimation [4], texture classification [5] and de-noising [6] for image or video applications. Although Loo and Kingsbury [7] had some works on DT-CWT watermarking, however none of their works dealt with the re-synchronization watermark embedding / detection; instead they use the original image to assist geometric invariant watermark detection.

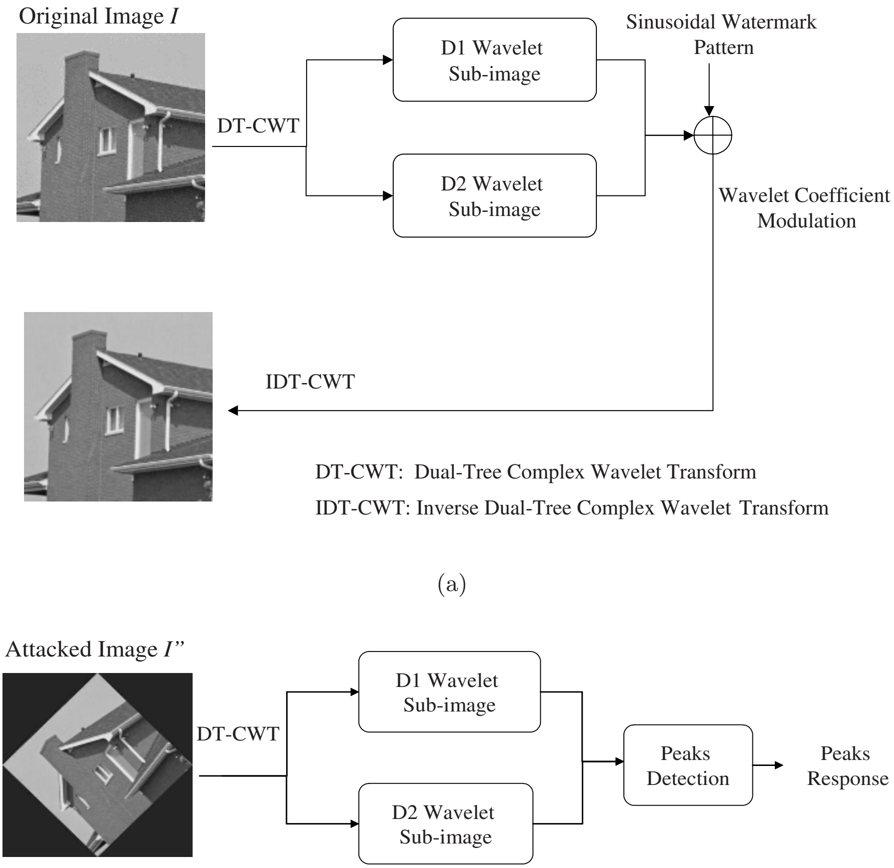
The goal of this paper is aiming at improving the robustness of re–synchronization peaks while preserving the picture quality of watermarked image. Section 2 discusses the proposed watermark re-synchronization embedding / detection algorithm. In section 3 experimental results are presented and analyzed. Finally, in section 4 conclusions as well as future work are given.

2 Proposed Algorithm

The main idea of the proposed scheme in Fig. 1(a) and (b) is to choose the robust DT-CWT sub-bands for embedding which will maximize the detection response peaks while preserving the visual quality of watermarked image at the same time. In the embedding process in Fig. 1(a), the original image first goes through 2-level DT-CWT decomposed into 14-subbands (Fig. 2). Sub-bands of each level are then grouped into two sub-images (D1: +15, +45 and +75, D2: -15, -45 and -75) based on their directionalities. In our preliminary implementation, only level-2 sub-images are considered for watermarking. The sub-bands of D1 and D2 are chosen to be modulated individually with the watermark pattern, and then are inverse-transformed to obtain the watermarked image. Depending on the respective detection robustness and the visual quality of watermarked image, one group of the sub-bands (either out of D1 or out of D2) will be eventually selected for embedding. In the detection process in Fig. 1(b), the attacked image first goes through 2-level DT-CWT decompositions, the sub-bands chosen out of D1 and D2 are then accumulated for peaks detection.

2.1 Sinusoidal Signals as Watermark Pattern

The watermark W consists of three sinusoidal signals, with frequency 16 cycles per unit length along x-axis, 32 cycles per unit length along y-axis, and 16/32 cycles per unit length along x/y axis. Fig. 3 shows the sinusoidal watermark



(a)

(b)

Fig. 1. (a) The flow of proposed watermark embedding scheme, (b) The flow of proposed watermark detection scheme.

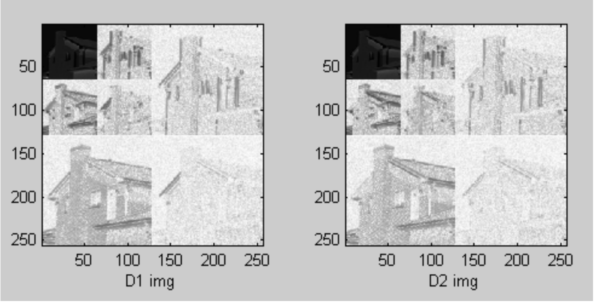
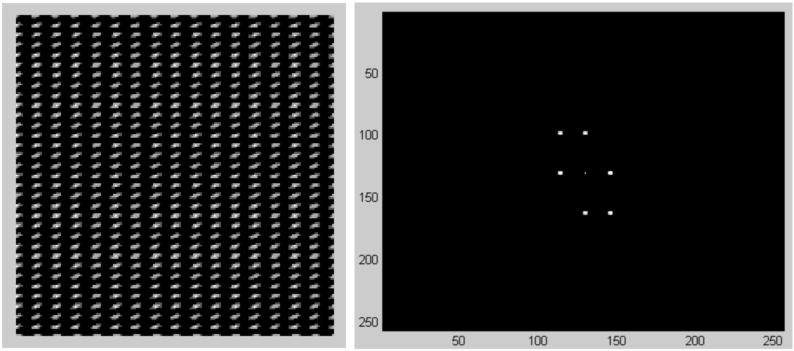


Fig. 2. 2-levels of DT-CWT decomposition



(a) (b)

Fig. 3. The sinusoidal watermark pattern and its peaks response in the frequency domain: (a) Sinusoidal watermark pattern (b) Peaks response

pattern and its peak response in the frequency domain. Ideally there should be totally 6 peaks (except for the center point) for these three sinusoidal signals, with 2 peaks for each sinusoid due to its symmetry in the frequency domain.

2.2 Watermark Embedding

The embedding algorithm consists of the following steps:

1) The original image *I* is decomposed into 14-subbands using the 2-level DT-CWT.

2) The sub-bands of each level are then grouped into two sub-images (*D1* and *D2*) based on their directionalities.

3) The sub-bands of *D1* and *D2* are chosen to be modulated individually with the watermark pattern *W*, and then are inverse-transformed to obtain the watermarked image *Iw'*.

4) Depending on their detection robustness and the visual quality of watermarked image, either sub-bands of *D1* or *D2* are eventually selected for embedding.

2.3 Watermark Detection

The detection algorithm consists of the following steps:

1) The attacked image *I"* (being attacked from *Iw'*) is decomposed into 14-subbands using the 2-level DT-CWT.

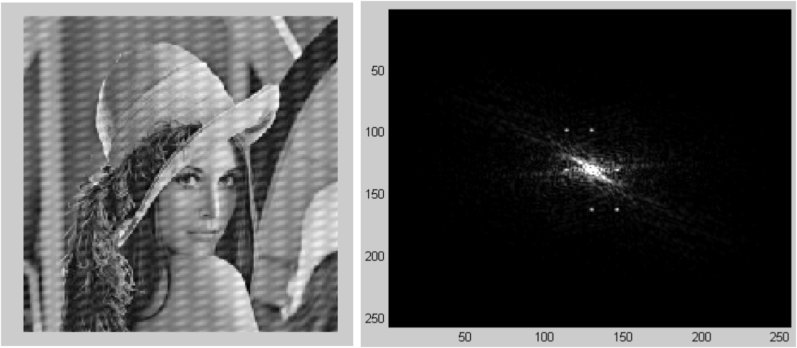
2) The sub-bands of each level are then grouped into two sub-images (*D1* and *D2*) based on their directionalities.

3) The sub-bands of *D1* and *D2* are then accumulated for peaks detection.

3 Experimental Results

To evaluate the effectiveness of the proposed method, three standard test images of size 256 x 256 including ”Lena”, ”Barbara” and ”Boat” are utilized as host signals to embed sinusoidal watermark patterns. To save space for the publication, only the results of ”Lena” image are shown here, other images show the similar results as well. Main issues regarding performance evaluation of the proposed method are discussed in the following.

The traditional spatial domain based scheme requires a very strong watermark embedding to ease re-synchronization peaks finding, which in turn leads to a poorly watermarked image. Figure 4 shows the watermarked image with PSNR 26.37dB and its corresponding peaks response. Note that the distortion caused by such strong embedding is quite visually significant.



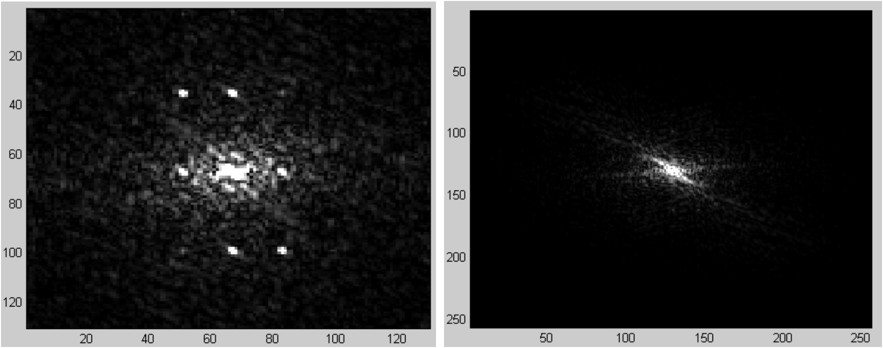
(a) (b)

Fig. 4. Watermarked image (PSNR 26.37dB) and its peaks response in the frequency domain of conventional scheme: (a) Watermarked image (b) Peaks response

For the fairness of comparison, the weighting parameter of watermark strength of our proposed approach is adapted to obtain the similar PSNR values (about 40dB) as that of the conventional spatial domain based scheme. Figure 5(a) and (b) show the watermarked image by adopting our proposed scheme and conventional one respectively, and their corresponding peak responses are shown in Fig. 6(a) and (b). We can see that the picture quality is very good for our scheme. However, there exists some small pattern artifact in the case of conventional scheme. The visibility of 6 peak responses is rather conspicuous for our proposed scheme. However, we cannot distinguish the 6 peak responses for

(a) (b)

Fig. 5. For the fairness of comparison, the watermarked images are adapted to obtain the similar PSNR values to test the robustness of peaks response: (a) Watermarked image of proposed scheme (PSNR 40.5 dB) (b) Watermarked image of conventional scheme (PSNR 40.3 dB)



(a) (b)

Fig. 6. The peak responses of our proposed scheme show far better effect than the approach of the conventional one: (a) Proposed scheme (b) Conventional scheme

conventional scheme. It is evident that our proposed scheme is superior to the approach of the conventional one.

To test the robustness of rotation, scaling and translation (RST), the watermarked image is manipulated with various combinations of RST transformations to generate attacked images. The peak responses are still visible under these various RST attacks of our proposed scheme. Figure 7 shows one of the image attacked by combinations of rotation 30°, scaling 1.8 and translation (20, 10). We can see from Fig. 8 that the 6 peak responses of our proposed scheme are rotated, scaled and translated by the same amounts as that of the attacked image. Our proposed scheme performs still far better than those of the conventional one.

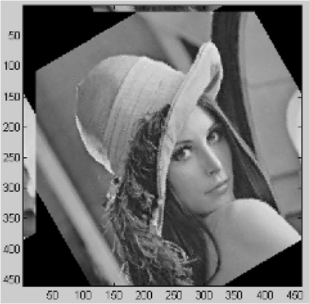
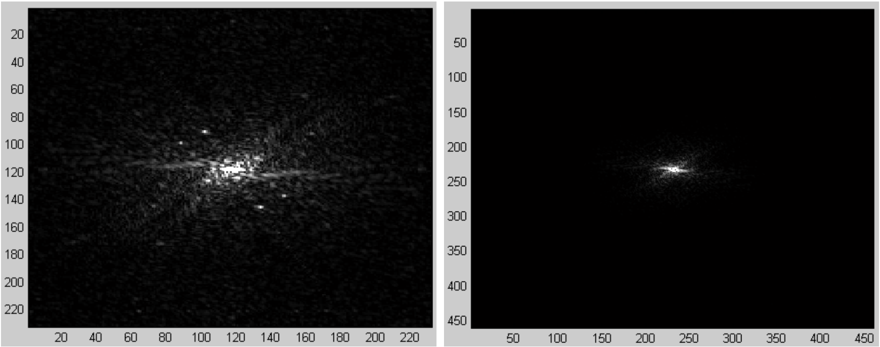


Fig. 7. Image attacked by combinations of rotation 30 ° , scaling 1.8 and translation (20, 10)



(a)

(b)

Fig. 8. In the case of attack by combinations of RST, the peak responses of our proposed scheme still perform far better than those of the conventional one: : (a) Proposed scheme (b) Conventional scheme

4 Conclusion

We propose in this paper a new approach for geometric invariant watermarking technique by embedding sinusoidal signals individually in each of the selected sub-bands of dual-tree complex wavelet domain (DT-CWT), and then detecting the re-synchronization peaks by using the accumulated embedding sub-bands. The main contributions of the proposed scheme are: (1) exploring the feasibility of taking DT-CWT as transform domain for watermark re-synchronization; (2) finding a viable choice for the robust DT-CWT sub-bands for embedding which will maximize the detection response peaks while preserving the visual quality of watermarked image. Experimental results demonstrate that by adopting our approach, the resulting re-synchronization peaks are robust to rotation, scaling and translation attacks while preserving the visual quality of the watermarked image, thereby resolve the unavoidable dilemma faced by the other schemes.

In the future, we expect to see more than 2-level decomposition of DT-CWT being employed, in which case the analysis of robust sub-bands for embedding will become increasingly complicated. On the other hand, further improvement could be made on the peaks finding algorithm to systematically find out the peaks location, so that real watermark payload (other than re-synchronization watermark patterns) embedding / detection could be achieved.

Acknowledgements. This work was supported in part by the National Science Council under grant 93-2219-E-009-018.

References

1. Martin Kutter: Watermarking resistance to translation, rotation, and scaling. Proc. SPIE Multimedia Systems Applications, Vol. 3528, 1998, pp. 423–432
2. David J. Fleet and David J. Heeger: Embedding invisible information in color images. In: Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438
3. Nick Kingsbury: Image Processing with complex wavelets. Phil. Trans. Royal Society London. A (19999) 357, 2543–2560
4. Julian Magarey and Nick Kingsbury: Motion Estimation using a Complex-Valued Wavelet Transform. IEEE Transactions on Signal Processing (April 1998) 1069–1084
5. Peter de Rivaz: Complex Wavelet Based Image Analysis and Synthesis. Ph.D Thesis, University of Cambridge (Oct 2000)
6. Levent Sendur and Ivan W. Selesnick: Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency. IEEE Transactions on Signal Processing (Nov 2002) 2744–2756
7. Patrick Loo: Digital Watermarking using Complex Wavelet. Ph.D Thesis, University of Cambridge (Mar 2002)

The Undeniable Multi-signature Scheme Suitable for Joint Copyright Protection on Digital Contents

Sung-Hyun Yun[1] and Hyung-Woo Lee[2]

[1] Div. of Information and Communication Engineering, Cheonan University,
Anseo-dong, Cheonan, 330-704, Korea
shyoon@cheonan.ac.kr

[2] Dept. of Software, Hanshin University, Osan, Gyunggi, 447-791, Korea
hwlee@hs.ac.kr

Abstract. In undeniable signature scheme, the digital signature can be verified and disavowed only with cooperation of the signer. Digital watermarks have been proposed as the means for copyright protection of multimedia data. Some one can provide counterfeited watermark which can be performed on a watermarked multimedia contents to allow multiple claims of rightful ownerships. The confirmer of a watermark wants to make sure that only the intended verifier can be convinced about the validity of the watermark. Undeniable signature scheme is suited to this application since the signature cannot be verified without cooperation of the signer. Existing copyright protection schemes are mainly focused on protection of single owner's copyright. In case that digital multimedia contents is made by co-workers, a joint copyright protection scheme is needed to provide equal right to them. In this study, we propose the undeniable multi-signature scheme. The multi-signature can be verified and disavowed only in cooperation with all signers. The proposed scheme satisfies the undeniable property and it is secure against active attacks such as modification and denial of the multi-signature by signers. We also discuss how to make joint copyright on digital contents and fair on-line sales of this copyrighted contents.

1 Introduction

An undeniable digital signature scheme is proposed by D.Chaum at first[3]. In the undeniable signature scheme, the digital signature can be verified and disavowed only with cooperation of the signer. There are many applications which an ordinary digital signature scheme can not be applied to[3,7,8]. A signed confidential document of a company can be copied and delivered to a rival company. If a conventional signature scheme is used to sign the document, it can be confirmed as authentic by verifying the signature without help of the signer. Therefore, if the company doesn't want the document to be verified as authentic by the rival company, it is recommended to use the undeniable signature scheme in which the signer can choose the verifier for signature confirmation.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 402–409, 2004.

A digital watermark is a signature added to digital data such that it could be used to uniquely establish ownership and to check if the data has been tampered with[9,10]. It has been proposed as the means for protection of copyright on multimedia contents. However, there are some problems in existing digital copyright protection mechanism as some one can provide counterfeited watermark which can be performed on a watermarked multimedia contents to allow multiple claims of rightful ownerships. In this case, the confirmer of a watermark wants to make sure that only the intended verifier can confirm the validity of the watermark.

The digital watermark generated by undeniable signature scheme differ from ordinary digital watermark in that the verifier can not distinguish whether this watermark is valid or not. Only the original watermarker can confirm it as authentic to the verifier.

Existing copyright protection mechanisms are mainly focused on protection of single author's copyright. Generally, the digital multimedia contents is made by many authors. In this case, joint copyright protection scheme is needed to provide equal right to them. In application that requires multiple authors and a designated verifier, an algorithm for undeniable multi-signature is needed.

In this study, the undeniable digital multi-signature scheme is proposed. The multi-signature can not be verified without cooperation of all signers. The proposed scheme satisfies undeniable property and it is secure against active attacks such as modification and denial of the multi-signature by signers. It also can be applied to protection of joint copyright on digital contents. In case of dispute between authors, the proposed scheme can resolve it by launching disavowal protocol to identify whether authors have cheated.

In section 2, the related work of our research is described. The proposed undeniable multi-signature scheme is presented in section 3 and undeniable property of our scheme is analyzed in section 4. In section 5, undeniable joint copyright protection scheme is discussed. Conclusion and future works are in section 6.

2 Related Works

In this section, we review existing signature schemes those concepts and properties are used to design the proposed scheme. The D.Chaum's undeniable signature scheme[3] and the El-Gamal signature scheme[2] are described. The security of these schemes are based on the difficulty of solving discrete logarithms over $GF(p)$[1, 2].

The D.Chaum's undeniable signature scheme consists of signature generation protocol, signature confirmation protocol and disavowal protocol. A Cryptographically secure galois field $GF(p)$ is defined in definition 1.

Definition 1 *If the following properties holds, it's computationally infeasible to solve discrete logarithms over galois field $GF(p)$.*
1) $p-1$ has large prime factor q.
2) G_q is the subgroup of $GF(p)$ having order q

2.1 D. Chaum's Undeniable Signature Scheme [3]

The signer randomly chooses private key x in Z_{q-1} and computes public key $y \equiv g^x \ (mod\ p)$. g is a generator of Gq.
(1) Signature generation protocol
Step 1: The signer computes the undeniable signature $z \equiv m^x \ (mod\ p)$ on the message m in G_q.
Step 2: The signer sends (m, z) to the verifier.
(2) Signature confirmation protocol
Step 1: The verifier chooses two random numbers (a, b) in Z_q and computes the challenge $w \equiv z^a y^b \ (mod\ p)$. The verifier sends w to the signer.
Step 2: The signer computes the response $R \equiv w^{x^{-1}} \ (mod\ p)$ and sends it to the verifier.
Step 3: The verifier computes $m^a g^b \ (mod\ p)$. If it is equal to R, then the signature z is valid. Otherwise, the following disavowal protocol is needed to identify whether the signature z is invalid or the signer has cheated.
(3) Disavowal protocol
The verifier chooses c, d in Z_q and computes the second challenge $w' \equiv z^c y^d \ (mod\ p)$. The signer computes the second response $R' \equiv w'^{x^{-1}} \ (mod\ p)$. Then the verifier computes the following discrimination equation.

$$(R \cdot g^{-b})^c \equiv (R' \cdot g^{-d})^a \ (mod\ p) : invalid\ signature$$

$$(R \cdot g^{-b})^c \not\equiv (R' \cdot g^{-d})^a \ (mod\ p) : denial\ by\ the\ signer$$

2.2 El-Gamal Digital Signature Scheme [2]

The signer chooses private key x in Z_{p-1} and computes public key $y \equiv g^x \ (mod\ p)$.
(1) Signature generation protocol
Step 1: The signer chooses a random number k in Z_{p-1}. k and $p-1$ are relatively prime integers.
Step 2: The signer computes the signature (r, s) on the message m as follows and sends (m, r, s) to the verifier.

$$r \equiv g^k \ (mod\ p), \quad m \equiv x \cdot r + k \cdot s \ (mod\ p-1), \quad m \in Z_{p-1}$$

(2) Signature confirmation protocol
If $g^m \equiv y^r \cdot r^s \ (mod\ p)$, the verifier confirms that the signature (r, s) is valid.

3 The Proposed Scheme

The proposed scheme consists of multi-signature generation protocol, multi-signature confirmation protocol and disavowal protocol. The multi-signature can not be verified without cooperation of all signers in the proposed scheme. The following parameters are used in the proposed scheme. Cryptographically secure

GF(p) is defined in definition 1.

Signers: $u_1, u_2, \ldots, u_n$ Message: m Hash function: h
Signer i's private key: $x_i \in Z_{p-1}, \quad 1 \le i \le n$,
Signer i's public key: $y_i \equiv g^{x_i}\ (mod\ p), \quad 1 \le i \le n$

3.1 Multi-signature Generation Protocol

The message drafter sends message m to all signers. Each signer computes undeniable signature and sends it to the message drafter. The message drafter uses each signer's undeniable signature to compute the undeniable multi-signature.
Step 1: The message drafter hashes the message m and sends (m, hpr) to the first signer u_1. The hpr is adjusted to make $m_h = h(m, hpr)$ as a primitive root of p.
Step 2: The u_1 chooses a random number k_1 and computes r_1 as follows. (r_1, Y_1) is delivered to the second signer u_2. Y_1 is used to make the common public key Y.

$$Y_1 = y_1, \quad r_1 \equiv {m_h}^{k_1}\ (mod\ p), \quad gcd(k_1, p-1) = 1, \quad k_1 \in Z_{p-1}$$

Step 3: The intermediate signer u_i $(2 \le i \le n)$ chooses a random number k_i and computes (r_i, Y_i) as follows.

$$r_i \equiv r_{i-1}^{k_i} \equiv m_h^{\prod_{j=1}^{i} k_j}\ (mod\ p), \quad Y_i \equiv Y_{i-1}^{x_i} \equiv g^{\prod_{j=1}^{i} x_j}\ (mod\ p)$$

The u_i sends (r_i, Y_i) to the next signer u_{i+1}. If the u_i is the last signer, (R, Y) is computed as follows. The last signer sends it to all signers as well as the message drafter.

$$R \equiv r_{n-1}^{k_n} \equiv {m_h}^{\prod_{j=1}^{n} k_j}\ (mod\ p), \quad Y \equiv Y_{n-1}^{x_n} \equiv g^{\prod_{j=1}^{n} x_j}\ (mod\ p)$$

Step 4: The signer u_i $(1 \le i \le n)$ computes the undeniable signature s_i and sends it to the message drafter. Since k_i and $p-1$ are relatively prime integers, there exists s_i satisfying the following equation.

$$k_i \cdot s_i \equiv x_i \cdot R - k_i \cdot m_h\ (mod\ p-1), \quad 1 \le i \le n$$

Step 5: The message drafter computes the undeniable multi-signature S as follows.

$$S \equiv \prod_{j=1}^{n} (m_h + s_j)\ (mod\ p)$$

3.2 Multi-signature Confirmation Protocol

In order to verify the multi-signature (R, S), the message drafter launches multi-signature confirmation protocol as follows.
Step 1: The message drafter chooses two random numbers (a, b) in Z_{p-1} and computes the challenge $ch \equiv R^{S \cdot a} \cdot Y^{R^n \cdot b}\ (mod\ p)$. The message drafter sends the challenge ch to the first signer u_1.

Step 2: The u_1 computes the response $rsp_1 \equiv ch^{x_1^{-1}}$ $(mod\ p)$ and sends it to the second signer u_2.
Step 3: The intermediate signer u_i $(2 \leq i \leq n)$ receives the response rsp_{i-1} from the signer u_{i-1}. Then the u_i computes the response $rsp_i \equiv rsp_{i-1}^{x_i^{-1}}$ $(mod\ p)$ and sends it to the next signer u_{i+1}. If the u_i is the last signer, the last signer sends the response rsp_n to the message drafter.
Step 4: The message drafter verifies the last signer's response rsp_n as follows.

$$rsp_n \equiv m_h^{R^n \cdot a} \cdot g^{R^n \cdot b}\ (mod\ p) \quad (3.1)$$

If equation 3.1 holds, the message drafter ensures that multi-signature is valid. Otherwise, the message drafter launches disavowal protocol to identify whether multi-signature is invalid or signers have cheated.

3.3 The Disavowal Protocol

The second challenge and response protocol launched by the message drafter is as same as that of the multi-signature confirmation protocol. The message drafter chooses two random numbers (c, d) and computes the second challenge $ch' \equiv R^{S \cdot c} \cdot Y^{R^n \cdot d}$ $(mod\ p)$. If the second response rsp_n' by the last signer is not equal to $m_h^{R^n \cdot c} \cdot g^{R^n \cdot d}$ $(mod\ p)$, additional step 5 is required.
Step 5: The message drafter makes the following discrimination equations. If R_1 equals to R_2, the message drafter confirms that multi-signature is invalid. Otherwise, at least more than one signer have cheated on valid multi-signature.

$$R_1 \equiv (rsp_n \cdot g^{-R^n \cdot b})^c\ (mod\ p), \quad R_2 \equiv (rsp_n' \cdot g^{-R^n \cdot d})^a\ (mod\ p)$$

4 Security Analysis

In this section, undeniable properties of the proposed scheme is analyzed. In theorem 1 and 2, we show correctness of our disavowal protocol.

Definition 2 *The valid multi-signature (R, S) and the invalid multi-signature (R', S) on the message m are defined as follows. X contains private keys of all signers.*

$$X \equiv \prod_{j=1}^{n} x_j\ (mod\ p-1), \quad m_h = h(m, hpr)$$

- *Valid multi-signature* (R, S)

$$R \equiv m_h^{\prod_{j=1}^n k_j}\ (mod\ p), \quad S \equiv \textstyle\prod_{j=1}^n (m_h + s_j)\ (mod\ p-1),$$
$$\textstyle\prod_{j=1}^n k_j(m_h + s_j) \equiv R^n \cdot X\ (mod\ p-1)$$

- *Invalid multi-signature* (R', S)

$$R' \equiv m_h^{\prod_{j=1}^n k_j{}'}\ (mod\ p), \quad \prod_{j=1}^{n} k_j'(m_h + s_j) \not\equiv R'^n \cdot X\ (mod\ p-1) \quad (4.1)$$

Equation 4.1 is for invalid multi-signature. We define X' which satisfying following equation, $\prod_{j=1}^{n} k'_j(m_h + s_j) \equiv R'^n \cdot X'$ $(mod\ p-1)$.

Theorem 1 *The proposed disavowal protocol can identify that the signers have compute the invalid response on the valid multi-signature.*

Proof: The valid multi-signature (R, S) is defined in definition 2. If more than one signer have cheated during the multi-signature confirmation protocol, X^{-1} which is used to compute the response is changed. We assume that X^{-1} is the valid inverse of X and X'^{-1} is the invalid inverse of X. The message drafter computes the challenge $ch \equiv R^{S \cdot a} \cdot Y^{R^n \cdot b}$ $(mod\ p)$. If more than one signer computes the response improperly on the valid multi-signature, the response rsp_n computed by all signers is written as follows.

$$rsp_n \equiv ch^{X'^{-1}} \equiv m_h^{a \cdot R^n \cdot X \cdot X'^{-1}} \cdot g^{b \cdot R^n \cdot X \cdot X'^{-1}}\ (mod\ p)$$

Since the rsp_n is not equal to $m_h^{R^n \cdot a} \cdot g^{R^n \cdot b}$ $(mod\ p)$, the message drafter launches disavowal protocol with new challenge value (c, d).

$$ch' \equiv R^{S \cdot c} \cdot Y^{R^n \cdot d}\ (mod\ p), \quad rsp'_n \equiv ch'^{X'^{-1}}\ (mod\ p)$$

The second response rsp'_n is not equal to $m_h^{R^n \cdot c} \cdot g^{R^n \cdot d}$ $(mod\ p)$. Therefore, the message drafter makes following discrimination equations.

$$R_1 \equiv (rsp_n \cdot g^{-R^n \cdot b})^c \equiv m_h^{a \cdot c \cdot R^n \cdot X \cdot X'^{-1}} \cdot g^{c \cdot (b \cdot R^n \cdot X \cdot X'^{-1} - b \cdot R^n)}\ (mod\ p)$$

$$R_2 \equiv (rsp'_n \cdot g^{-R^n \cdot d})^a \equiv m_h^{c \cdot a \cdot R^n \cdot X \cdot X'^{-1}} \cdot g^{a \cdot (d \cdot R^n \cdot X \cdot X'^{-1} - d \cdot R^n)}\ (mod\ p)$$

From the above equations, R_1 is not equal to R_2. Therefore, we prove the correctness of the proposed disavowal protocol in case that more than one signer have cheated on the valid multi-signature. Q.E.D.

Theorem 2 *The proposed disavowal protocol can identify that the multi-signature is invalid.*

Proof: The invalid multi-signature (R', S) is defined in definition 2. The first challenge and response on (R', S) is as follows.

$$ch \equiv R'^{S \cdot a} \cdot Y^{R'^n \cdot b}\ (mod\ p),\ rsp_n \equiv m_h^{a \cdot R'^n \cdot X' \cdot X^{-1}} \cdot g^{b \cdot R'^n}\ (mod\ p)$$

The second challenge and the response on (R', S) is as follows.

$$ch' \equiv R'^{S \cdot c} \cdot Y^{R'^n \cdot d}\ (mod\ p),\ rsp'_n \equiv m_h^{c \cdot R'^n \cdot X' \cdot X^{-1}} \cdot g^{d \cdot R'^n}\ (mod\ p)$$

The message drafter makes the following discrimination equations. In the following equation, R_1 is equal to R_2.

$$R_1 \equiv (rsp_n \cdot g^{-R'^n \cdot b})^c \equiv m_h^{a \cdot c \cdot R'^n \cdot X \cdot X'^{-1}}\ (mod\ p)$$

$$R_2 \equiv (rsp'_n \cdot g^{-R'^n \cdot d})^a \equiv m_h^{c \cdot a \cdot R'^n \cdot X \cdot X'^{-1}}\ (mod\ p)$$

Therefore, we prove the correctness of the proposed disavowal protocol in case that the multi-signature is invalid. Q.E.D.

5 Undeniable Joint Copyright Protection

In this section, we show that the proposed scheme is suitable for protecting copyright on digital contents made by several authors where all authors want to share copyright together.

5.1 Undeniable Property of Digital Watermark

Digital watermarks are verified as authentic by anyone using the verification process. However, this *self-verification* property is unsuitable from many applications such as copyrighted commercial multimedia contents distribution. The validity or invalidity of an undeniable digital watermark can be ascertained by conducting a verification process with original author. If a confirmation process is needed, the cooperating original author gives exponentially-high certainty to the verifier that the digital watermark does correspond to the legal one.

We can use undeniable signature scheme in the watermarking process. Multimedia contents company could embed the watermark which they signed using an undeniable signature scheme. Only someone who had directly purchased the contents from that company could verify the watermark and be certain that the contents were right. However, if the company sold contents which were not right, they should be unable to deny their watermark afterwards.

5.2 Protecting Joint Copyright on Digital Contents

Many authors can participate jointly in authoring of digital multimedia contents. In this case, the copyright of the digital contents must be shared by all participants. The proposed scheme can be used to protect joint copyright on the digital contents as follows. Joint-copyright on it is created in the multi-signature generation protocol. Each author computes the undeniable signature and sends it to the copyright maker. The copyright maker computes the undeniable multi-signature and watermarks it to the digital contents. For sales on digital contents by on-line, a customer can buy it by launching multi-signature confirmation protocol. Without the consent of all authors, the customer can not buy digital contents. Especially, in case of dispute between authors, the proposed disavowal protocol can discriminate whether authors have cheated or the digital watermark is invalid.

6 Conclusion

In this paper, we propose the undeniable digital multi-signature scheme. The undeniable property of the proposed scheme is proved and its application for joint copyright protection is discussed. If a watermarking scheme used in embedding joint copyright can not be recovered from attack such as bit distortion errors, the proposed copyright protection mechanism also can not correctly recover its joint copyright data. In this case, the proposed scheme only can identify these copyrights as invalid ones. In the streaming protocol based contents sales model like live internet video/audio service, there are some bit errors to maintain good quality of real-time service on multimedia contents. Our scheme is not suited to this application. The proposed scheme is best suited to the sales model where complete downloadable multimedia contents is correctly distributed to the customers by trusted on-line transaction. It also can be used in applications where fragile watermarking schemes are required.

Acknowledgments. This research was supported by Brain informatics Research Program sponsored by Korean Ministry of Science and Technology.

References

1. W.Diffie, M.E.Hellman, "New Directions in Cryptography," IEEE Transactions on Information Theory, Vol. IT-22, No. 6, pp. 644–654, 1976
2. T.Elgamal, "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms," IEEE Transactions on Information Theory, Vol. IT-31, No. 4, pp. 469–472, 1985
3. D.Chaum, "Undeniable Signatures," Advances in Cryptology, Proceedings of CRYPTO'89, Springer-Verlag, pp. 212–216, 1990
4. L.Harn, "(t,n) Threshold Signature and Digital Multisignature," Workshop on Cryptography & Data Security, pp. 61–73, 1993
5. M.Naor, "Bit Commitment using Pseudorandomness," In Advances in Cryptology, Proceedings of CRYPTO'89, LNCS 435, pp. 128–136, 1990
6. S.H.Yun, T.Y.Kim, "A Digital Multisignature Scheme Suitable for EDI Message," Proceedings of 11th International Conference on Information Networking, pp. 9B3.1–9B3.6., 1997
7. S.H.Yun, T.Y.Kim, "Convertible Undeniable Signature Scheme," Proceedings of IEEE High Performance Computing ASIA'97, pp. 700–703, 1997
8. S.H.Yun, S.J.Lee, "An electronic voting scheme based on undeniable signature scheme," Proceedings of IEEE 37th carnahan conference on Security Technology, pp. 163–167, 2003
9. Andre Adelsbach, Birgit Pfitzmann, Ahmad-Reza Sadeghi : Proving Ownership of Digital Content, 3rd International Information Hiding Workshop (IHW '99), LNCS 1768, Springer-Verlag, 117–133, 1999
10. Andre Adelsbach, Ahmad-Reza Sadeghi: Zero-Knowledge Watermark Detection and Proof of Ownership, 4th International Information Hiding Workshop (IHW '01), LNCS 2137, Springer-Verlag, 273–288, 2001

A Digital Watermarking Scheme for Personal Image Authentication Using Eigenface

Chien-Hsung Chen and Long-Wen Chang

Institute of Information Systems and Applications
National Tsing Hua University
101 Sections 2, Kung Fu Road
Hsinchu, Taiwan, 30055
lchang@cs.nthu.edu.tw

Abstract. Recently, face recognition has received a great of attention because digital camera can be widely available. It has many applications including security management in the airport and the IC card for e-commerce. In this paper, we introduce a watermark technique for image authentication with a personal face. We try to compute the eigenface of a personal image and encode the eigenface features with bar-code. The bar-coded feature values are then embedded into the image that is to be authenticated. Our simulation shows the similarity measures of the watermark extracted from the damaged image under various attack are vary high and indicate the proposed method is very feasible for real applications.

Keywords: Digital watermarking, eigenface, personal authentication

1 Introduction

The internet technology has become more and more popular and convenient. In the internet, one can easily get almost all information they need, and easily access a remote computer. Therefore, the network security is very important [1][2]. Recently, people pay lots of attention to the copyright protection [4]-[6]. The copyright protection plays an important role in e-commerce [7]. In this paper, we show a watermark system for personal image copyright protection. We try to use the biometric information [8][9] and extract "face feature value" [11] and then treat it as the watermark. We combine it with the bar-code mechanism [12]-[14]. Then, we embed the barcoded eigenface feature data in the image. In our experiment it is practical and seems to be useful in real applications.

2 The Proposed Biometric Watermarking System

The biometric watermarking system includes 3 parts: Embedding, Extraction, and Verifying procedures. In the embedding procedure we first process the owners (or the buyers) face image. We calculate its feature value (The weight distribution which project on the face space) and transform it to bar-code watermark.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 410–417, 2004.

In embedding time, we transform only parts of feature value to watermark. We use these parts of feature value as the "face key", then we embed it to target image. In the verifying procedure, we use owner's (or buyer's) face image and calculate its feature values. We compare these feature values with the values extracted from extraction procedure. If feature values are the same or the similarity is beyond some threshold, we treat the image owner as a legal user. If feature values are different, the owner is an illegal user.

2.1 The Embedding Procedure

In the embedding procedure, first we should collect a bunch of face images. These images are the training set; we use them to form a face space. Before we use these face images, we must "crop" them so that the eyes and chin are included. Other useless information will be discarded. After we select and process these face images, we transform them to the eigenface space. The flow chart of the embedding procedure shows in Fig. 1.

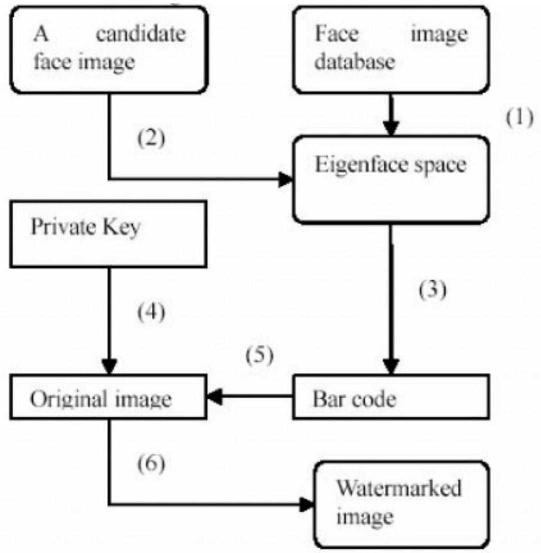


Fig. 1. The embedding procedure

(1) Compute the eigenvectors of them and sort these eigenvectors according to its eigenvalues. We keep only M eigenfaces with the largest eigenvalues.

(2) Take owner or buyer's face image and compute the distribution weight on the eigenface space. We calculate these eigenvalues according to the specific eigen-face space, so we will get different eigenvalues on every different eigenface space. In our system, we must choose a set of face space first and should remember it for later use (The verifying procedure will need to know the face space so we can compute the eigenvalues).

(3) Encode the eigenvalue information into a bar-code image. We don't use the known bar code standard and try to extend the bar pattern so we can encode full ASCII in the bar code. The reason we use bar-code is that it is a meaningful graphic, and has good recovery ability. Once our watermark suffers attacks, we can easily repair it. In fact, in our system we only use bar-code to encode the eigenvalue information and the eigenvalue information can be represent by numbers ($0 \sim 9$) and "lower case" characters ($a \sim z$). So we only need to encode 36 characters in our bar-code system.

(4) Take owner's or buyer's private key and use it in the SHA-1 algorithm. SHA-1 is a one way hash function to generate different sequences with different keys. We use this sequence to generate 20 different watermark embedding positions so we won't embed watermark in the fixed position. This step will improve security.

(5) Embed the bar-code image in the frequency domain (DCT domain). The positions we embed bar-code were generated in step 4.

(6) We get the watermarked image.

2.2 The Extraction Procedure

In the extraction procedure, we get the embedded bar-code image for later use (In verifying procedure). The flow chart of the extraction procedure shows in Fig. 2.

1. We use original image ($I_{512\times512}$) to map with the watermarked image (${I'}_{512\times512}$).

If $I'_i \leq I \rightarrow W_i = 0$; Else if $I'_i > I \rightarrow W_i = 255$

Use above equation, we can construct the bar-code image.

2. We compute the watermark embedding position and thus can construct the bar-code image

3. After extracting and constructing the bar-code image, we may see that bar-code image has some "noise" signal.

4. We try to repair the bar-code image. The method is not difficult and the rules we use are listed as below:

1. We can determine the color of the line is black or white by count the black pixel numbers (B_i) and the white pixel numbers (W_i) in that line (i_{th} line).

If $B_i > W_i \rightarrow$ The color of the i_{th} line is black

Elseif $B_i < W_i \rightarrow$ The color of the i_{th} line is white

In this step, we also record the number of error pixels in that line. The error pixel information can be used in step 3 to differentiate bold lines and thin lines.

2. If two neighbouring lines have the same color. We can treat them as a bold line. We can check the number of bold lines and thin lines to see if it fit to our bar-code spec.

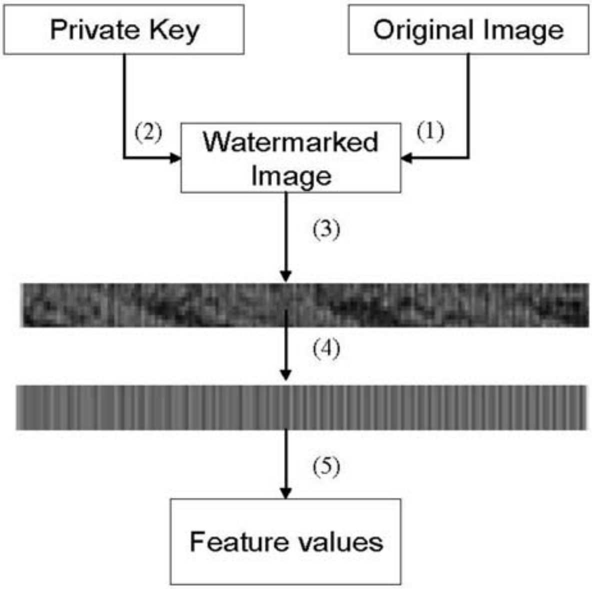


Fig. 2. The extraction procedure

3. If the number of bold lines and thin lines is not correct, we can try to change the color of the line which has maximum error pixels until it satisfies the bar-code spec.

4. We decode bar-code image and get the feature sequence information. The feature sequence information will be used in our verifying procedure.

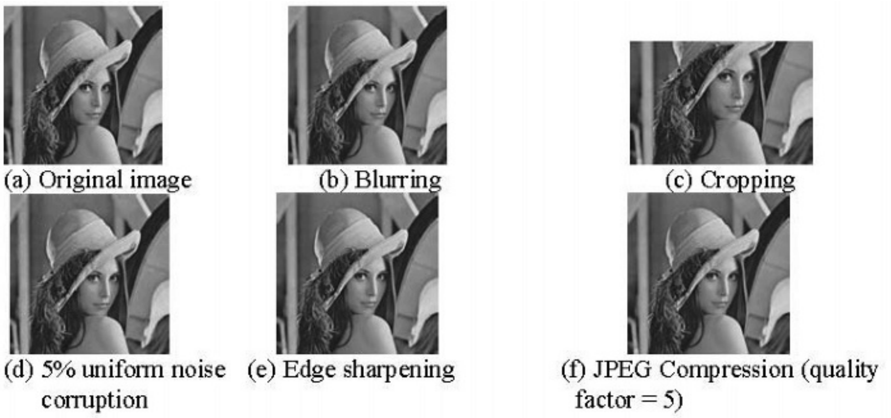
3 Experimental Results

In our experiment, we use 512*512 "Lena" image to be the test image. First we embed bar-code watermark in the image, then we try some image processing attack on the watermarked image. The image processing attack we use include "Blur", "Noise", "Jpeg compression", "Crop", and "Sharpen". After these attacks, we compute PSNR to see the image quality. We also compute the error character number between the extracted feature value and the original feature value. The equation we used to compute the similarity is listed as below:

$$Sim = 1 - \frac{E_n}{N},$$

where En means error character number, N is the length of feature values Table 1 is the private key we use to compute the watermark embedding location.

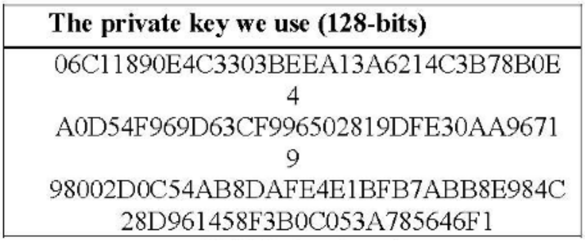
We select 15 face images from "yalefaces". One of the images is used as the "owner or buyer's face image" and the other 14 images form the eigenspace. We


(a) Original image (b) Blurring (c) Cropping

(d) 5% uniform noise corruption (e) Edge sharpening (f) JPEG Compression (quality factor = 5)

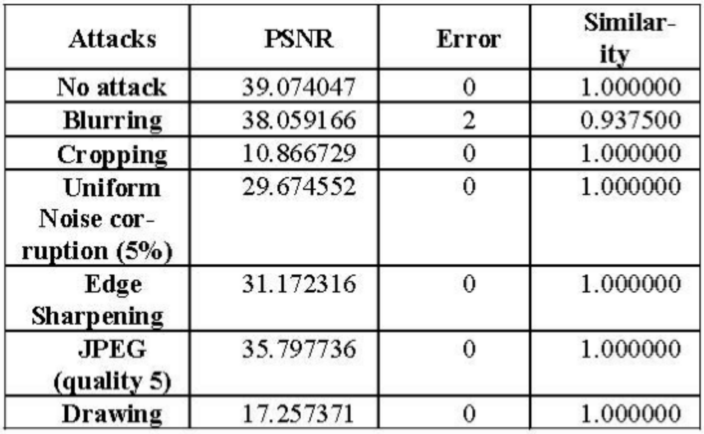
Fig. 3. The watermarked images under various attacks

Table 1. A private key



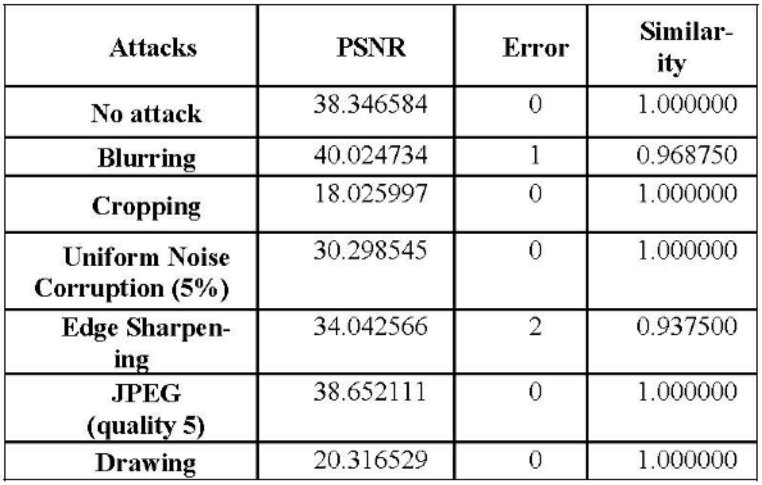
The private key we use (128-bits)
06C11890E4C3303BEEA13A6214C3B78B0E 4 A0D54F969D63CF996502819DFE30AA9671 9 98002D0C54AB8DAFE4E1BFB7ABB8E984C 28D961458F3B0C053A785646F1

Table 2. Similarity measures for the "Lena" image after various attacks



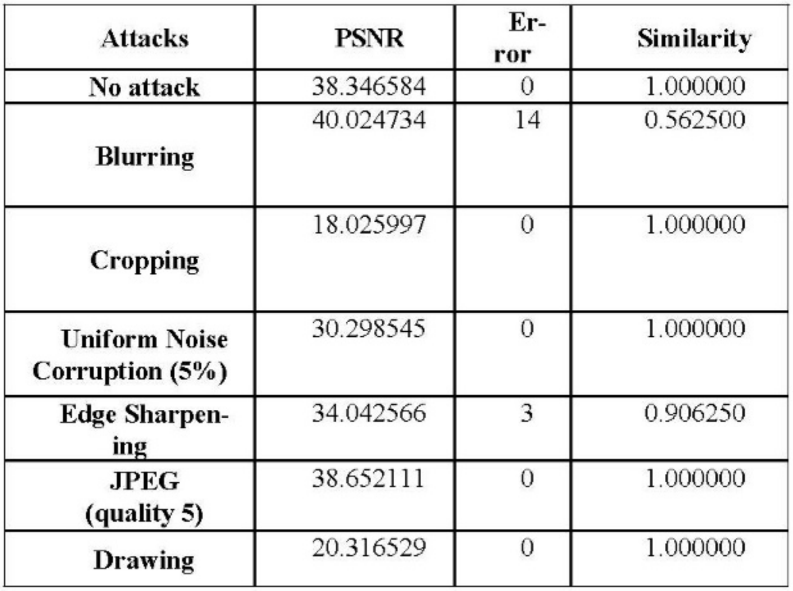
Attacks	PSNR	Error	Similar-ity
No attack	39.074047	0	1.000000
Blurring	38.059166	2	0.937500
Cropping	10.866729	0	1.000000
Uniform Noise cor-ruption (5%)	29.674552	0	1.000000
Edge Sharpening	31.172316	0	1.000000
JPEG (quality 5)	35.797736	0	1.000000
Drawing	17.257371	0	1.000000

Table 3. Similarity measures for the "Jet" image after various attacks



Attacks	PSNR	Error	Similarity
No attack	38.346584	0	1.000000
Blurring	40.024734	1	0.968750
Cropping	18.025997	0	1.000000
Uniform Noise Corruption (5%)	30.298545	0	1.000000
Edge Sharpening	34.042566	2	0.937500
JPEG (quality 5)	38.652111	0	1.000000
Drawing	20.316529	0	1.000000

Table 4. Similarity measures for the "Barbara" image after various attacks



Attacks	PSNR	Error	Similarity
No attack	38.346584	0	1.000000
Blurring	40.024734	14	0.562500
Cropping	18.025997	0	1.000000
Uniform Noise Corruption (5%)	30.298545	0	1.000000
Edge Sharpening	34.042566	3	0.906250
JPEG (quality 5)	38.652111	0	1.000000
Drawing	20.316529	0	1.000000

select 14 face images as the image space and then compute their eigenvalues. We only keep the larg-est eigenvalue as the eigenspace and then calculate I_o's distribution weight on it. The bar-code image we used in this experiment can encode 64 characters. It is enough to encode 32-char long feature values here. After encoding the feature value and embedding it in the original image, we try some attacks and see the result. Figure 3 shows the watermarked image under various attacks. The result is shown in Table 2. Besides "Lena" image, we also use "Jet" and "Barbara" image for test. The image processing attacks which process on these test images are the same. Table 3, 4 shows the results for Jet and Barbara images. All these image processing attacks are processed by "photoshop". We can see that after most attacks, bar-code can recover itself. However, after some attacks like "blur" and "sharpen", the extracted watermark still have some error character. These kinds of attacks may cause uniform changes in bar-code image. After recovery, most watermark messages' similarity is beyond the threshold 0.9.

4 Conclusion

In our watermark system, we use eigenface feature value as the "face key" and combine it with the bar-code mechanism for personal image authentication. Bar-code is the combination of regular line patterns, and the damaged bar-code can be recovered well in most situations. We use eigenface feature value information as the watermark and use SHA-1 hash function to select the watermark embedding position. The benefits are that third party can't extract the watermark unless they have the private key we use. Our simulation shows that the proposed method is very feasible for real applications.

References

1. W. Stallings: "Cryptography and Network Security: Principles and Practice", 2nd edn, Prentice Hall, (1999)
2. Bruce Schneier: "Applied Cryptography", Second Edition, John Wiley and Sons, (1996)
3. N. Nikolaidis and I. Pitas: "Copyright protection of images using robust digital signatures," in Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 4, (May 1996) 2168–2171
4. S. Craver, N. Memon, B. L and M. M Yeung: "Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications," IEEE Journal on Selected Areas in Communications, Vol. 16 Issue: 4, (May 1998) 573–586
5. Siang-Gen Sia, Charles G. Boncelet, and Gonzalo R. Arce: "A Multi-resolution Watermark for Digital Image" in IEEE (1997) 518–521
6. I. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure Spread Spectrum Wa-termarking for Multimedia", IEEE Transaction on Image Processing, vol. 6, (Dec.1997) 1673–1687
7. Nasir Memon, Ping Wah Wong: "A Buyer-Seller Watermarking Protocol", IEEE Transactions on Image Processing, Vol. 10, No. 4, (April 2001) 643–649
8. Anil K. Jain, Umut Uludag and Rein-Lien Hsu: "Hiding a face in a fingerprint image", Proc. ICPR 2002, 16. International Conference on Pattern Recognition, vol. 3, pp. 756–759, Quebec City, Quebec, Canada, August 11-15, (2002)
9. Anil K. Jain, Umut Uludag: "Hiding fingerprint minutiae in images", Proc. AutoID 2002, 3. Workshop on Automatic Identification Advanced Technologies, pp. 97–102, Tarrytown, New York, USA, March 14-15, 2002 (2002)

10. Matthew Turk, Alex Pentland: “Eigenfaces for Recognition”, Journal of Cognitive Neutroscience Vol 3, No. 1, (1991)
11. Bruce Schneier: “Inside risks: the uses and abuses of biometrics”, Source Communications of the ACM, Vol. 42 , Issue 8, pp. 136, (August Dec.1999)
12. Hiroo Wakaumi, Takatoshi Komaoka, Eiji Hankui: “Grooved Bar-Code Recognition System with Tape-Automated-Bonding Head Detection Scanner”, IEEE Transactions on Magnetics, Vol. 36, No. 1, pp. 366–370, (January 2000)
13. Kuroki, M.; Yoneoka, T.; Satou, T.; Takagi, Y.; Kitamura, T.; Kayamori, N.: “Barcode recognition system using image processing”, Emerging Technologies and Factory Automation Proceedings, 1997. ETFA ’97, 1997 6th International Conference on, 9-12 (Sept. 1997) 568–572
14. Wakaumi, H.; Ajiki, H.; Hankui, E.; Nagasawa, C.: “Magnetic grooved bar-code recognition system with slant MR sensor”, Science, Measurement and Technology, IEE Proceedings, Volume: 147 Issue: 3, (May 2000) 131–136

Cryptanalysis of a Chaotic Neural Network Based Multimedia Encryption Scheme

Chengqing Li[1], Shujun Li[2a⋆], Dan Zhang[3], and Guanrong Chen[2b]

[1] Department of Mathematics, Zhejiang University,
Hangzhou 310027, Zhejiang, China, swiftsheep@hotmail.com
[2] Department of Electronic Engineering, City University of Hong Kong,
Kowloon Toon, Hong Kong, China, hooklee@mail.com[2a], eegchen@cityu.edu.hk[2b]
[3] College of Computer Science, Zhejiang University,
Hangzhou 310027, Zhejiang, China, zhangdan@etang.com

Abstract. Recently, Yen and Guo proposed a chaotic neural network (CNN) for signal encryption, which was suggested as a solution for protection of digital images and videos. The present paper evaluates the security of this CNN-based encryption scheme, and points out that it is not secure from the cryptographical point of view: 1) it can be easily broken by known/chosen-plaintext attacks; 2) its security against the brute-force attack was much over-estimated. Some experiments are shown to support the results given in this paper. It is also discussed how to improve the encryption scheme.

1 Introduction

In the digital world today, the security of multimedia data (such as digital speeches, images, and videos) becomes more and more important since the communications of such digital signals over open networks occur more and more frequently. Also, special and reliable security in storage and transmission of multimedia products is needed in many real applications, such as pay-TV, medical imaging systems, military image/database communications and confidential video conferences, etc. To fulfill such a need, many encryption schemes have been proposed as possible solutions [1, Sec. 4.3], among which some are based on chaotic systems [1, Sec. 4.4]. Meanwhile, cryptanalysis work has also been developed, which reveal that some proposed multimedia encryption schemes have been known to be insecure.

From 1998, Yen et al. proposed a number of chaos-based multimedia encryption schemes [1, Sec. 4.4.3], but some of them have been successfully broken by Li et al. [2,3,4,5,6]. This paper analyzes the security of a class of encryption schemes proposed by Yen et al. in [7,8,9], which have not yet been cryptanalyzed before. In a recent paper [10], this class of encryption schemes were simply extended to arbitrary block size without influencing the security and then applied for JPEG2000 image encryption.

[⋆] The corresponding author, web site: http://www.hooklee.com.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 418–425, 2004.

The studied encryption scheme here is a stream cipher based on a chaotic neural network (CNN), which is designed to encrypt 1-D signals and is simply extended to encrypt 2-D digital images and 3-D videos. This paper evaluates the security of the CNN-based scheme and points out two security problems: 1) it can be easily broken by the known/chosen-plaintext attacks with only one known/chosen plaintext; 2) its security against the brute-force attack was much over-estimated.

The rest of the present paper is organized as follows. In Sec. 2, a brief introduction of the CNN-based encryption scheme is given. The cryptanalytic studies and some experimental results are given in Sec. 3. Section 4 briefly discusses how to improve the security of the studied encryption scheme, and the last section concludes the paper.

2 The CNN-Based Scheme for Signal Encryption

In the following, the concerned encryption scheme is simply referred to as CNN.

Assuming that $\{f(n)\}_{n=0}^{M-1}$ is a 1-D signal for encryption, the encryption procedure of CNN can be briefly depicted as follows:

- The *chaotic Logistic map* $f(x) = \mu x(1-x)$ is used, where μ is the control parameter [11].
- *The secret key* is the control parameter μ and the initial point $x(0)$ of the Logistic map, which are all L-bit binary decimals.
- *The initialization procedure*: under L-bit finite computing precision, run the Logistic map from $x(0)$ to get a chaotic sequence $\{x(i)\}_{i=0}^{\lceil 8M/K\rceil-1}$, and extract K bits below the decimal dot of each chaotic state[1] to generate a chaotic bit sequences $\{b(i)\}_{i=0}^{8M-1}$, where $x(i) = 0.b(Ki+0)\cdots b(Ki+K-1)\cdots$.
- *The encryption procedure*: For the n-th plain-element $f(n) = \sum_{i=0}^{7} d_i(n)\times 2^i$, the corresponding cipher-element $f'(n) = \sum_{i=0}^{7} d'_i(n)\times 2^i$ is determined by the following process:
  - for $i = 0 \sim 7$ and $j = 0 \sim 7$, 64 weights w_{ji} are calculated as follows: if $i = j$, $w_{ji} = 0$; else $w_{ji} = 1 - 2b(8n+i) = \begin{cases} 1, & b(8n+i) = 0, \\ -1, & b(8n+i) = 1; \end{cases}$
  - for $i = 0 \sim 7$, 8 biases θ_i are calculated as follows:

$$\theta_i = \frac{2b(8n+i)-1}{2} = \begin{cases} -1/2, & b(8n+i) = 0, \\ 1/2, & b(8n+i) = 1; \end{cases}$$

  - the i-th cipher-bit $d'_i(n)$ is calculated as follows:

$$d'_i(n) = \operatorname{sign}\left(\sum\nolimits_{j=0}^{7} w_{ji}\times d_i(n) + \theta_i\right), \tag{1}$$

[1] In real implementations of CNN, the K bits can be extracted from the direct multiplication result $\mu x(i-1)(1-x(i-1))$, before $x(i)$ is obtained by quantizing the value. As a result, it is possible that $K > L$. For example, in [9], $K = 32 > L = 17$.

where sign$(\cdot)$ denotes the sign function, i.e., $\text{sign}(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$

- *The decryption procedure* is the same as the above one.

The above encryption procedure looks very complicated, however, actually it can be simplified to be a much more precise form. Observing the proofs of Proposition 1 in [7,8] and Lemma 1 in [9], one can see the following fact:

$$d_i'(n) = \begin{cases} 0, & \text{if } d_i(n) = 0 \text{ and } b(8n+i) = 0, \\ 1, & \text{if } d_i(n) = 1 \text{ and } b(8n+i) = 0, \\ 1, & \text{if } d_i(n) = 0 \text{ and } b(8n+i) = 1, \\ 0, & \text{if } d_i(n) = 1 \text{ and } b(8n+i) = 1, \end{cases} \tag{2}$$

which means that

$$d_i'(n) = d_i(n) \oplus b(8n+i), \tag{3}$$

where $\oplus$ denotes the XOR operation.

Obviously, CNN is a stream cipher encrypting the plain-signal bit by bit, where the key stream for masking is the chaotic bit sequence $\{b(i)\}$.

3 Cryptanalysis of the CNN-Based Encryption Scheme

3.1 Brute-Force Attacks

In [7,8,9], it was claimed that the computing complexity of a brute-force attack to CNN is $O\left(2^{8M}\right)$, since there are $8M$ bits in $\{b(i)\}_{i=0}^{8M-1}$ (which is unknown to the attacker). However, this statement is not true due to the following fact: the $8M$ bits are uniquely determined by the secret key, i.e., the control parameter μ and the initial condition $x(0)$, which have only $2L$ secret bits. This means that there are only 2^{2L} different chaotic bit sequences.

Now, let us see what is the real complexity of a brute-force attack. For each guessed value of $x(0)$ and μ, about $8M/K$ chaotic iterations and $8M$ XOR operations are needed for verification. Assuming that each L-bit digital multiplication needs L times of additions, then each chaotic iteration needs $2L+1$ times of additions. Therefore, the complexity of a brute-force attack to CNN will be $O\left(2^{2L} \times \left(\frac{8M(2L+1)}{K} + 8M\right)\right) = O\left(2^{2L}M\right)$, which is much smaller than 2^{8M} when M is not too small. What's more, considering the fact that the Logistic map can exhibit strong chaotic behavior only when μ is close to 4 [11], the complexity should be even smaller than $O\left(2^{2L}M\right)$.

The above analysis shows that the security of CNN was much over-estimated by the authors, even under the simplest attack. Because of the rapid progress of digital computer and distributed computing techniques, the complexity not lower than $O\left(2^{128}\right)$ is required for a cryptographically strong cipher [12]. To achieve such a security level, $L \geq 64$ is required. As a comparison, $L = 8$ in [8] and $L = 17$ in [9], which are both too small[2].

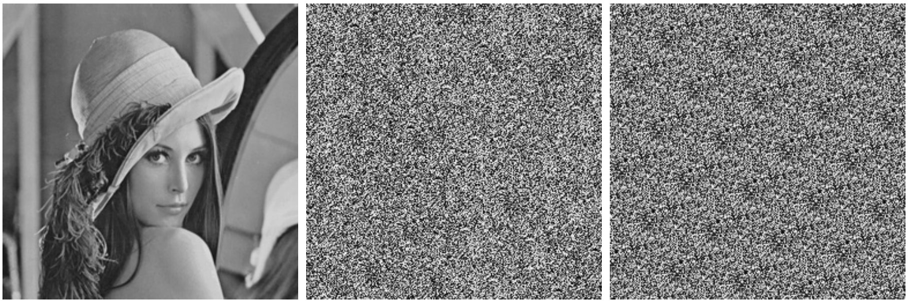
[2] In [7], the value of L is not explicitly mentioned. Since [7] is an initial version of [8], it is reasonable to assume $L = 8$.

3.2 Known/Chosen-Plaintext Attacks

In known-plaintext or chosen-plaintext attacking scenarios, CNN can be broken with only one known/chosen plaintext $\{f(n)\}_{n=0}^{M-1}$ and its corresponding ciphertext $\{f'(n)\}_{n=0}^{M-1}$, with a complexity that is smaller than the complexity of a brute-force attack.

From Eq. (3), one can get $b(8n+i) = g_i(n) \oplus g_i'(n)$. That is, an attacker can successfully reconstruct the chaotic bit sequence $\{b(i)\}_{i=0}^{8M-1}$ by simply XORing $\{f(n)\}_{n=0}^{M-1}$ and $\{f'(n)\}_{n=0}^{M-1}$ bit by bit. Assuming $\{f_m(n) = f(n) \oplus f'(n)\}_{n=0}^{M-1}$, one has $f_m(n) = 0.b(8n+0)\cdots b(8n+7)$. Without deriving the secret key $(\mu, x(0))$, given any ciphertext g' encrypted with the same secret key, the attacker can use f_m to decrypt the M leading bytes of the corresponding plaintext g: $n = 0 \sim M-1$, $g(n) = g'(n) \oplus f_m(n)$. Here, we call f_m the *mask signal* (or the *mask image* when CNN is used to encrypt digital images), since the plaintext can be decrypted by using f_m to "mask" (i.e., XOR) the ciphertext[3].

To demonstrate the above attack, with the parameters $L = 17, K = 32$ [9] and the secret key $\mu = 3.946869$, $x(0) = 0.256966$, some experiments are given for the encryption of digital images. In Fig. 1, a 256 × 256 known/chosen plain-image "Lenna", its corresponding cipher-image, and the mask image $f_m = f \oplus f'$ are shown. If another plain-image "Babarra" (of size 256×256) is encrypted with the same key, it can be broken with the mask image f_m derived from "Lenna" as shown in Fig. 2. For a larger plain-image "Peppers" (of size 384×384), the 256×256 leading pixels can be successfully broken with f_m as shown in Fig. 3.

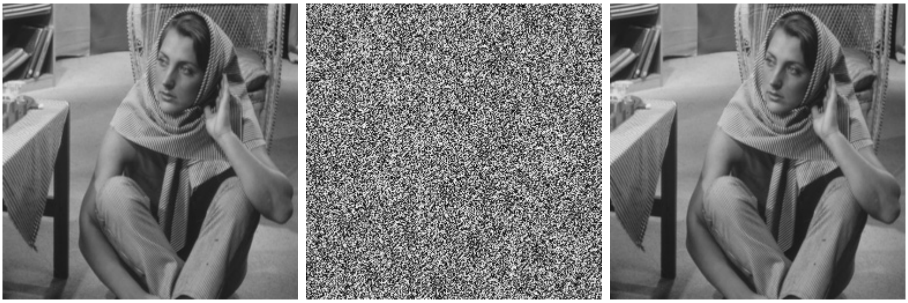


a) The plain-image f b) The cipher-image f' c) The mask image f_m

Fig. 1. One known/chosen plain-image "Lenna" (256×256), its corresponding cipher-image, and the mask image $f_m = f \oplus f'$

From the above experiments, one can see that the breaking performance of known/chosen-plaintext attacks based on f_m is limited. Fortunately, from the reconstructed bit sequence $\{b(i)\}_{i=0}^{8M-1}$, it is easy for an attacker to derive the values of μ and $x(0)$, and then to completely break CNN. Even when only part of a plaintext $f(n_1) \sim f(n_2)$ is known to the attacker, he can still derive the

[3] In fact, it is a common defect of most stream ciphers [12].

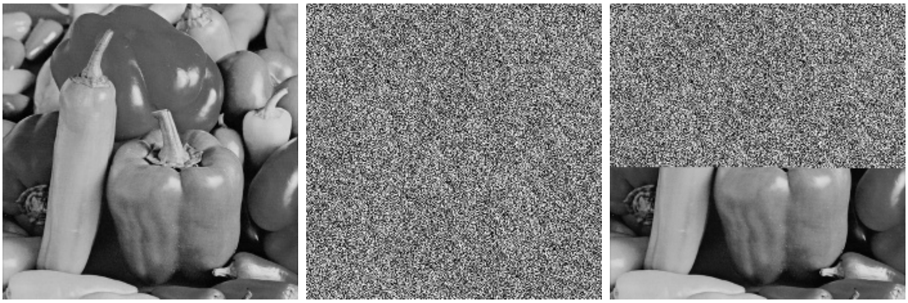


a) The plain-image "Babarra"

b) The encrypted "Babarra"

c) The recovered "Babarra" with f_m

Fig. 2. Decrypt a plain-image "Babarra" (256 × 256) with f_m shown in Fig. 1c



a) The plain-image "Peppers"

b) The encrypted "Peppers"

c) The recovered "Peppers" with f_m

Fig. 3. Decrypt a plain-image "Peppers" (384 × 384) with f_m shown in Fig. 1c

values of μ and a chaotic state $x(i)$, which can be used to calculate all following chaotic states, i.e., all following chaotic bits $\{b(i)\}_{i=8n_2}^{\infty}$. In this case, all plain-pixels after the n_1-th position can be broken. In the following, let us discuss how to derive chaotic states and the value of μ.

Firstly, let us see how a chaotic state $x(i)$ is derived. Recall the generation procedure of $\{b(i)\}_{i=0}^{8M-1}$. It is easy to reconstruct a K-bit approximate of the chaotic sequence by dividing $\{b(i)\}_{i=0}^{8M-1}$ into K-bit segments: $\{\widetilde{x}(i)\}_{i=0}^{\lceil 8M/K\rceil-1}$, where $\widetilde{x}(i) = 0.b(Ki+0)\cdots b(Ki+K-1)$ and

$$|\varDelta x(i)| = |\widetilde{x}(i) - x(i)| \leq 0.\overbrace{0\cdots 0}^{K}\overbrace{1\cdots 1}^{L-K} = \sum_{j=K+1}^{L} 2^{-j} < 2^{-K}. \qquad (4)$$

Apparently, when $L \leq K$, $\widetilde{x}(i) = x(i)$; when $L > K$, the exact value of each chaotic state $x(i)$ can be derived by exhaustively guessing the $L-K$ unknown bits, and the guess complexity is $O\left(2^{L-K}\right)$.

Once two consecutive chaotic states $x(i)$ and $x(i+1)$ are derived, the estimated value of μ can be calculated to be $\tilde{\mu} = \frac{x(i+1)}{x(i)\cdot(1-x(i))}$. Due to the influence of quantization errors existing in forward chaotic iterations, in general $\tilde{\mu} \neq \mu$. When the difference between $\tilde{\mu}$ and μ is sufficiently small, it is possible to exhaustively search the neighborhood of $\tilde{\mu}$ to find the accurate value of μ with a sufficiently small complexity. In the following, we will show how to get a $\tilde{\mu}$ close enough to μ, and estimate the search complexity of the accurate value of μ.

Apparently, the estimation error $\Delta\mu = \tilde{\mu} - \mu$ is caused by the quantization error $\Delta x(i+1)$ generated in the forward chaotic iteration $x(i+1) = \mu \cdot x(i) \cdot (1 - x(i))$. In one L-bit digital multiplication, the quantization error does not exceed 2^{-L} for the floor or ceiling quantization function, and does not exceed $2^{-(L+1)}$ for the round quantization function. Considering there are two L-bit digital multiplications in each forward chaotic iteration, one has

$$\begin{aligned}\bar{x}(i+1) &= (\mu \cdot x(i) + \Delta_1 x(i+1)) \cdot (1 - x(i)) + \Delta_2 x(i+1)\\ &= \mu \cdot x(i) \cdot (1 - x(i)) + \Delta_1 x(i+1) \cdot (1 - x(i)) + \Delta_2 x(i+1)\\ &= x(i+1) + \Delta x(i+1),\end{aligned}$$

where $\bar{x}(i+1)$ denotes the real value of $x(i+1)$ and $\Delta x(i+1) = \Delta_1 x(i+1) \cdot (1 - x(i)) + \Delta_2 x(i+1)$. Then, one can get $|\Delta x(i+1)| \leq |\Delta_1 x(i+1)| + |\Delta_2 x(i+1)| < 2^{-L} + 2^{-L} = 2^{-(L-1)}$, and get the quantization error $|\Delta\mu|$ as follows:

$$\begin{aligned}|\Delta\mu| &= \left|\frac{\Delta x(i+1)}{x(i) \cdot (1 - x(i))}\right| = \left|\frac{\Delta x(i+1)}{x(i+1)} \cdot \frac{x(i+1)}{x(i) \cdot (1 - x(i))}\right|\\ &= \frac{|\Delta x(i+1)|}{x(i+1)} \cdot \mu < \frac{2^{-(L-1)}}{x(i+1)} \cdot 4 = \frac{1}{2^{L-3} \cdot x(i+1)}.\end{aligned} \tag{5}$$

When $x(i+1) \geq 2^{-n}$ ($n = 1 \sim L$), $|\Delta\mu| < \frac{1}{2^{L-3} \cdot x(i+1)} \leq \frac{1}{2^{L-3} \cdot 2^{-n}} = 2^{n+3} \times 2^{-L}$, which means the size of the neighborhood of $\tilde{\mu}$ for exhaustive search is 2^{n+3}. To minimize the search complexity in real attacks, $x(i+1) \geq 0.5$ is suggested to derive μ, which occurs with a probability of 0.5. In this case, $n = 1$ and the size of the searched neighborhood is only $2^{3+1} = 16$.

With the mask image f_m derived from the known plain-image "Lenna" (of size 256×256) shown in Fig. 1a, the values of $x(0)$ and μ are calculated following the above procedure to completely decrypt the larger plain-image "Peppers" (of size 384×384). The decryption result is given in Fig. 4.

Finally, it deserves being mentioned that even without deriving the secret key there is another way based on a mask signal f_m to decrypt any plaintext of arbitrary size. It is due to the following fact: for a digital chaotic system implemented in L-bit finite computing precision, each chaotic orbit will lead to a cycle whose length is smaller than 2^L (and generally much smaller than 2^L, see [4, Sec. 2.5]). For the implementation of CNN in [9], $L = 17$, $K = 32$. Thus, the cycle length of each chaotic orbit will be much smaller than 2^{17} in most cases. Such a length is not sufficiently large in comparison with the size of many plaintexts, especially for digital images and videos. For example, a 256×256 image corresponds to a chaotic orbit $\{x(i)\}$ whose length is $8 \times 256 \times 256/32 = 2^{14}$. For almost every value of μ and $x(0)$, the cycle length of $\{x(i)\}$ is even much smaller than 2^{14}, which means that there exists an visible repeated pattern in

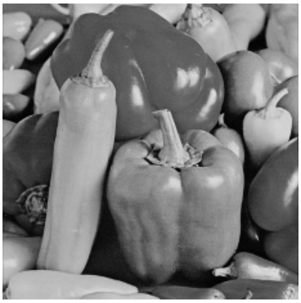
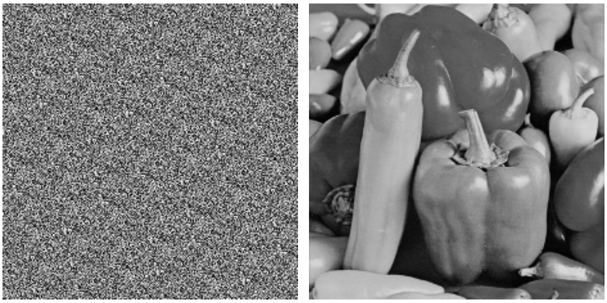


Fig. 4. The decrypted "Peppers" (384 × 384) with the secret key derived from f_m shown in Fig. 1c



a) The extended mask image f_m^*

b) The recovered "Peppers" with f_m^*

Fig. 5. Decrypt "Peppers" (384 × 384) with f_m^* extended from f_m shown in Fig. 1c

$\{x(i)\}$. Carefully observing the mask image f_m shown in Fig. 1c, one can easily find such a repeated pattern. Then, it is easy to get the cycle of f_m, and to extend it to arbitrary sizes by appending more cycles at the end of the original mask signal. This means that any ciphertext can be decrypted with a mask signal f_m^* extended from the mask image f_m. Using such a method, the larger plain-image "Peppers" is completely decrypted as shown in Fig. 5.

4 Improving the CNN-Based Encryption Scheme

The simplest way to improve the original CNN is to make L sufficiently large so as to ensure the complexity of the brute-force attack cryptographically large. In addition, to make the complexity of guessing the $L - K$ unknown bits of each chaotic state cryptographically large, $L - K$ should also be sufficiently large. To be practical, $(L, K) = (64, 8)$ is suggested. In this case, the complexity to get the value of $x(0)$ is $O\left(2^{L-K}\right) = O\left(2^{56}\right)$, and the complexity to get the value of μ (i.e., to get two consecutive chaotic states) is $O\left(2^{2(L-K)}\right) = O\left(2^{112}\right)$. Such a complexity is sufficiently large to make both the brute-force attack and the attack of deriving the secret key from f_m impossible in practice.

However, because CNN is a stream cipher, making $L - K$ sufficiently large cannot enhance the security against the known/chosen-plaintext attacks based on the mask signal f_m. To resist such attacks, a substitution encryption part should be used to make CNN a product cipher. Note that the security of the modified CNN is ensured by the new substitution part, not the CNN itself. So, essentially speaking, the CNN cannot be enhanced to resist known/chosen-plaintext attacks.

5 Conclusion

In this paper, the security of a chaotic signal encryption scheme called CNN [7, 8,9,10] has been investigated and it is found that the encryption scheme is not secure from the cryptographical point of view. Both theoretical and experimental analyses show the feasibility of the proposed known/chosen-plaintext attacks of breaking CNN. Also, it is pointed out that the security of CNN against brute-force attacks was much over-estimated. Some possible methods to enhance the security of CNN are also discussed, but its insecurity against the known/chosen-plaintext attacks cannot be essentially improved. As a result, CNN is not suggested in applications requiring a high level of security.

Acknowledgements. This research was supported by the National Natural Science Foundation, China, under grant no. 60202002, and by the Applied R&D Centers of the City University of Hong Kong under grants no. 9410011 and no. 9620004.

References

1. Li, S., Chen, G., Zheng, X.: Chaos-based encryption for digital images and videos. In Furht, B., Kirovski, D., eds.: Multimedia Security Handbook, Chapter 4. CRC Press (2004) The preprint is available at `http://www.hooklee.com/pub.html`.
2. Li, S., Zheng, X.: Cryptanalysis of a chaotic image encryption method. In: Proc. IEEE Int. Symposium on Circuits and Systems. Volume II. (2002) 708–711
3. Li, S., Zheng, X.: On the security of an image encryption method. In: Proc. IEEE Int. Conference on Image Processing. Volume 2. (2002) 925–928
4. Li, S.: Analyses and New Designs of Digital Chaotic Ciphers. PhD thesis, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China (2003) Available online at `http://www.hooklee.com/pub.html`.
5. Li, S., Li, C., Chen, G., Mou, X.: Cryptanalysis of the RCES/RSES image encryption scheme. Submitted to IEEE Trans. Image Processing (2004)
6. Li, S., Li, C., Chen, G., Zhang, D., Bourbakis, N.G.: A general cryptanalysis of permutation-only multimedia encryption algorithms. In preparation (2004)
7. Yen, J.C., Guo, J.I.: A chaotic neural network for signal encryption/decryption and its VLSI architecture. In: Proc. 10th VLSI Design/CAD Symposium. (1999) 319–322
8. Su, S., Lin, A., Yen, J.C.: Design and realization of a new chaotic neural encryption/decryption network. In: Proc. IEEE Asia-Pacific Conference on Circuits and Systems. (2000) 335–338
9. Yen, J.C., Guo, J.I.: The design and realization of a chaotic neural signal security system. Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications) **12** (2002) 70–79
10. Lian, S., Chen, G., Cheung, A., Wang, Z.: A chaotic-neural-network-based encryption algorithm for JPEG2000 encoded images. In: Advances in Neural Networks - ISNN 2004: International Symposium on Neural Networks Proceedings, Part II. Lecture Notes in Computer Science **3174** (2004) 627–632
11. Hao Bai-Lin: Starting with Parabolas: An Introduction to Chaotic Dynamics. Shanghai Scientific and Technological Education Publishing House, Shanghai, China (1993) (in Chinese).
12. Schneier, B.: Applied Cryptography – Protocols, Algorithms, and Souce Code in C. Second edn. John Wiley & Sons, Inc., New York (1996)

A Secure Steganographic Method on Wavelet Domain of Palette-Based Images[⋆]

Wei Ding, Xiang-Wei Kong, Xin-Gang You, and Zi-Ren Wang

School of Electronic and Information Engineering,
Dalian University of Technology, China
weiding_dlut@hotmail.com, {kongxw,youxg}@dlut.edu.cn

Abstract. This article presents a novel secure steganographic method on wavelet domain of GIF images. Secret information is usually embedded in palettes or indices of GIF images directly by formerly presented steganographic methods. These methods may introduce visible noise and detectable changes of parameters in images. The new method based on integer wavelet transform dispels noise introduced by data-hidding into adjacent pixels. Matrix encoding is also applied in embedding. Both scattering noise and matrix encoding improve the quality of the stego-images and the security of secret communication. Experimental results show the fine security of the proposed method in resisting attacks by χ^2 detecting method and Fridrich's detecting method.

Keywords: Steganography; GIF image; integer wavelet; matrix encoding; security

1 Introduction

GIF images are popular carriers used in steganography because of its widely use on Internet. Although steganographic algorithms on GIF images appeared early, there are not so many secure methods at present for many detecting methods were also presented. GIF format contains a palette and image indices pointing to the corresponding colors in the palette. There are 256 kinds of colors at most in the palette. Most of current steganographic methods embed secret information directly in palettes or indices of GIF Images. There are mainly three classes of approaches. Gif-shuffle [1] uses different combinations of colors to embed secret message, leading to a limited capacity of 210 bytes. Artifacts are easy to be detected by χ^2 steganalysis method [2] in stego-images created with softwares such as S-Tools [3] or Hide&Seek [4] which change the palette and image indices simultaneously. Schemes changing indices directly such as EZ Stego [5] and methods presented by Fridrich [6,7,8] may introduce visible noise which deteriorates the quality of images. And EZ Stego can not counteract the steganalysis methods presented by Fridrich [9].

[⋆] This paper is supported by the project of National Nature Science Foundational of China (Grant No. 60372083).

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 426–432, 2004.

In this article, a novel secure steganographic algorithm is presented which embeds secret information in the high frequency coefficients of the wavelet domain. Thus, noise generated by embedding message is scattered into adjacent pixels. We implement matrix encoding in the process of embedding which also improves the security of the algorithm. The principle of the new approach is introduced in section 2. Section 3 lists experimental results that show the security of the algorithm in defending χ^2 detecting method and steganalysis method presented by Fridrich.

2 Method Description

2.1 Integer Wavelet Transform

In GIF images, image indices are stored as integer and its scale is 0∼255 when there are 256 kinds of colors in the palette. If transform in common use such as DCT or DWT is used in steganography [10] on GIF images, the overflow in spacial domain is hard to control. So we should adopt the integer wavelet transform. S transform presented by Swelden [11] is adopted in this paper. It is the Harr integer wavelet transform mapping integers to integers. The overflow can be controlled by preprocessing that we will present in section 2.3. The S transform is:

$$l = \left\lfloor \frac{x+y}{2} \right\rfloor, \quad h = y - x \tag{1}$$

The inverse transform is:

$$x = l - \left\lfloor \frac{h}{2} \right\rfloor, \quad y = l + \left\lfloor \frac{h+1}{2} \right\rfloor \tag{2}$$

When we apply these formulas to images, x and y denote the adjacent pixels values in a row or a volume of the image. l and h are the low frequency and high frequency part respectively. "$\lfloor\rfloor$" means "the greatest integer less than or equal to".

2.2 Matrix Encoding

Matrix encoding[12,13] is used to improve the security of the algorithm. When the length of secret message is less than the maximum capacity of the cover image, the number of changes due to message embedding can be decreased by adopting matrix encoding. When a secret message x with k bits is going to be hidden in a code a which contains n modifiable positions, we can find a proper code a' using matrix encoding for code a. Let f be a hash function that extracts k bits from code a. The hash function $f(a)$ can be determined by the followed equation:

$$f(a) = \bigoplus_{i=1}^{n} a_i * i \tag{3}$$

where $\oplus$ represents the operation of exclusive or and a_i is the ith modifiable position in the code a. The code a' is the modified code a which is generated by the f function and message x with $x = f(a')$. The Hamming distance $d(a, a')$ follows:

$$d\left(a, a'\right) \leq d_{\max} \tag{4}$$

In the formula $d_{\max}$ represents the maximum of changed indices without matrix encoding when message is embedded under the same condition. Thus we use $(d_{\max}, n, k)$ to represent this matrix encoding. The discussion presented by Westfield[13] shows that the embedding efficiency using matrix encoding is higher than that without matrix encoding.

2.3 Algorithm of the New Method

The indices scale is 0∼255 for GIF images with 256 colors. Embedding data in coefficients obtained from integer wavelet transform directly may overflow the range of image index value. The overflow will result in the failure of extracting information. The preprocessing adopted by this method to overcome this difficulty is described as follows.

Preprocessing. Let x' , y' be the indices of the stego-image and x, y are the indices of the corresponding pixel in the cover image (0≤x≤255,0≤y≤255). Δx, Δy are the values of the modification in spatial domain, thus:

$$x' = x + \Delta x, \quad y' = y + \Delta y \tag{5}$$

Let Δh be the value of the modification of the high frequency coefficients after embedding, then from (2):

$$x' = l - \left\lfloor \frac{h + \Delta h}{2} \right\rfloor, \quad y' = l + \left\lfloor \frac{h + 1 + \Delta h}{2} \right\rfloor \tag{6}$$

From (5) and (6) we can obtain the following:

$$\Delta x = \left\lfloor \frac{h}{2} \right\rfloor - \left\lfloor \frac{h + \Delta h}{2} \right\rfloor, \quad \Delta y = \left\lfloor \frac{h + 1 + \Delta h}{2} \right\rfloor - \left\lfloor \frac{h + 1}{2} \right\rfloor \tag{7}$$

According to the deduction above we can find the changing direction of the scale. Δh may be equal to –1, 0 or 1 during embedding. The changed value of the index may be –1, 0 or 1 after inverse transform according to (7). Then the range of indices changes from (0∼255) to (-1∼256). So the scale of indices in the cover should be adjusted to (1,254) in preprocessor in order to keep the indices of the stego image in the normal range. The process is presented as follows:

1. Let $f(c_i)$ be the appearance frequency of color c_i; Let A and B be the two indices of colors which have the least appearance frequency in the palette:

$$\begin{aligned} f(A) &= \min f(c_i) \ (i = 0, 1, \cdots, 255) \\ f(B) &= \min f(c_i) \ (i = 0, 1, \cdots, 255, i \neq A) \end{aligned} \tag{8}$$

2. C and D are the closest color indices to A and B respectively according to Euclidean norm distance in the palette (C≠A≠B, D≠A≠B).
3. Let *width* and *height* be the width and height of the image respectively. Let $index_j$ be the index of the jth pixel in the image. Replace A, B with C,D respectively in image indices:

$$index_j = \begin{cases} C \quad if \quad index_j = A \\ D \quad if \; index_j = B \quad 0 \le j < width * height \\ index_j \quad otherwise \end{cases} \tag{9}$$

4. Replace 254, 255 with A and B, respectively in the original palette. Indices should be adjusted again because of the change of the palette:

$$index_j = \begin{cases} A \quad if \quad index_j = 254 \\ B \quad if \; index_j = 255 \quad 0 \le j < width * height \\ index_j \quad otherwise \end{cases} \tag{10}$$

At last preprocessor is finished.

Create the New Palette. The palette should be rearranged after the preprocessor. The arrangement algorithm is similar to traveling salesman problem. It is based on the principle that the sum of the Euclidean norm distance between rearranged adjacent colors is the least among all kinds of orders. So reordered neighboring color indices are close to each other. The process to create the new palette is described as Fig. 1. The first 254 colors rearranged. The indices range of those 254 colors in the new palette is assigned to (1,254). Then the first position in the new palette is filled with the second color and the last position is filled with its previous color. The original palette is also required to adjust to correspond the new palette. The procedure is described as follows:

OldPal[254]= NewPal[1] OldPal[255]= NewPal[254]

In the above description newPal[i] and oldPal[i] represent the ith position in the new palette and the original palette respectively.

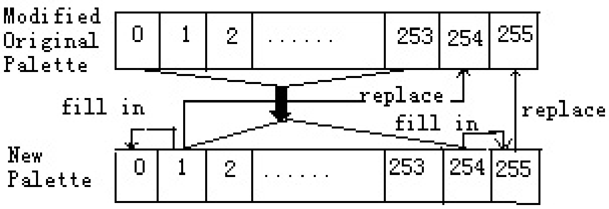


Fig. 1. producing the new palette

Procedure for Embedding. New image indices can be obtained when the new palette is created. Then embedding process with matrix encoding (1, n, k) is presented as following:

1. Let *maxcap* be the maximum of the embedding capacity of the image:

$$maxcap = width * height * 0.5 \tag{11}$$

2. Let *msglg* be the actual length of the secret message. Then proper n and k could be counted according to the formulas:

$$\begin{aligned} msglg > (maxcap * (k+1)/n - (maxcap * (k+1)/n\%n)) \\ msglg \leq (maxcap * k/n - maxcap * k/n\%n) \end{aligned} \tag{12}$$

   The operation '%' is to get the integral remainders.
3. Code a is composed by the LSBs of the n high frequency coefficients selected according to the order generated by a pseudo random seed. The hash function $f(a)$ can be determined by (3). When k bits message x is hidden in code a, the position s to be changed[13] is gained by:

$$s = x \oplus f(a) \tag{13}$$

   At last the modified code a' is obtained by:

$$a^{'} = \begin{cases} a, \quad if\ s = 0 (\Leftrightarrow s = f(a)) \\ (a_1, a_2, \cdots, \neg a_s, \cdots, a_n), \quad otherwise \end{cases} \tag{14}$$

   where ¬ is the bit-wise *not* operation.
4. Repeat step 3 to embed the next k bits until all message is embedded. Then apply inverse wavelet transform according (2).Store image data as GIF format.

Extract Secret information. Secret information could be extracted according to the inverse procedure of embedding.

3 Experimental Results and Discussion

Stego images created by the proposed method and EZ stego are called new-stego-images and EZ-stego-images respectively. Experimental results of PSNR shown in Table 1 indicate the good quality of new-stego-imges. In this section, χ^2 detecting[2] method and Fridrich's detecting method[9] are used to test the security of the new algorithm. The embedding capacity in experiments is 0.5bit/pixel. Experimental images are from different sources which contain 60 sheets of scanned images, 65 sheets of images from digital camera and 20 sheets of images from Fabien's standard image database on Internet[14]. As Table 1 shows new-stego-images counteract χ^2 steganalysis method for the detecting rate is

Table 1. Experimental Results of χ^2 steganalysis method and the average PSNR. The threshold is 60%; If the embedding probability counted from a image by χ^2 steganalysis method is above 60%, the image is considered as a stego-image

Image Class	Scanned Images	Digital Photos	Standard Images
Image Quantity	60 sheets	65 sheets	20 sheets
Detecting Rate	6.19%	4.16%	0%
Average PSNR	32.3098 dB	34.5622 dB	34.6144 dB

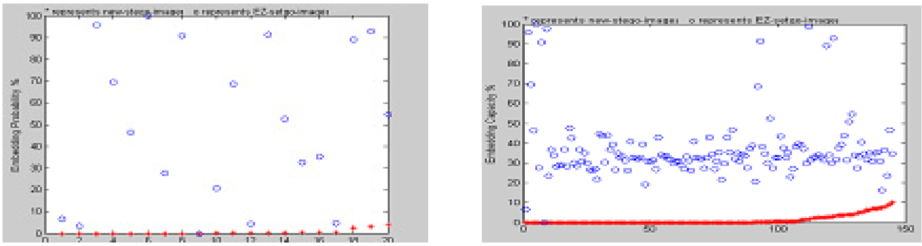


Fig. 2. Comparison of χ^2 detecting results on 20 sheets of images from Fabien's standard image database(the left) and camparison of Fridrich detecting results on 145 sheets of images(the right); 100% represents the embedding of capacity of 1bit/pixel; $*$ represents new-stego-image;

very low. Figure 2 show the comparison between new-stego-images and EZ-stego-images using χ^2 detecting method(the left) and Fridrich detecting method(the right). The comparisons indicate high security of the new algorithm. The security of new-stego-images is better than that of EZ-stego-images because that histogram of the image after embedding data by EZ Stego has many adjacent pairs but this phenomena in the new-stego-images is unconspicuous and noises in new-stego-images introduced by steganography is scattered into adjacent indices. All these experiments make it clear that the new method has fine performance.

4 Conclusion and Outlook

Experimental results show that new-stego-images produced by the new method has good quality and this method has high performance in defending χ^2 steganalysis method and steganalysis method presented by Fridrich. But artifacts might be visible in some GIF Images after embedding because they have small number of colors. So in futrue work, we should take more attention on the improvement of visual security in images with small number of colors.

References

1. Kwan, M.: Gifshuffle 2.0. Available from http://www.darkside.com.au/gifshuffle/ IEEE Trans. (2003)
2. A. Westfeld and A. Pfitzmann: Attacks on Steganographic Systems. Lecture Notes in Computer Science,vol.1768, Springer-Verlag, Berlin,(1999) pp. 61–76
3. Brown A.: S-Tools for Windows, Shareware.
www.jjct.com/steganography/toolmatrix.htm (1994)
4. Colin Moroney : Available from www.jjct.com/steganography/toolmatrix.htm
5. Machado, R: EZ Stego, Stego Online, Stego. Available from http://www.stego.com (1997)
6. Fridrich, J.: Applications of data hiding in digital images. Tutorial for The ISSPA_99, Brisbane, Australia, (1999)
7. Fridrich, J.: A new steganographic method for palette-based images. IS&T PICS, Savannah, Georgia, 25-28 (1999) pp. 285–289
8. Fridrich, J., Du, J.: Secure steganographic methods for palette images. In: Proc. the Third Inform. Hiding Workshop LNCS, vol. 1768. Springer-Verlag, New York, pp. 47–60
9. Jessica Fridrich, Miroslav Goljan, David Soukal: Higher-order statistical steganalysis of palette images. In Proc. EI SPIE Santa Clara, CA, Jan (2003) pp. 178–190
10. Rufeng Chu, Xingang You, Xiangwei Kong, Xiaohui Ba: A DCT-based Image Steganographic Method Resisting Statistical Attacks. The 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2004)
11. Calderbank A R, Daubechies I, Sweldens W, et al: Wavelet transforms that map integers to integers[R]. Princeton, New Jersey, U.S.: Department of Mathematics, Princeton Universitv. (1996)
12. Ron Crandall: Some Notes on Steganography. Posted on Steganography Mailing List. http://os.inf.tu-dresden.de/.westfeld/crandall.pdf (1998)
13. Andreas Westfeld.: F5—A Steganographic Algorithm High Capacity Despite Better Stega nalysis. IH 2001, LNCS 2137, (2001) pp. 289–302
14. http://www.petitcolas.net/fabien/watermarking/image_database

Mapping Energy Video Watermarking Algorithm Based on Compressed Domain

Lijun Wang, HongXun Yao, ShaoHui Liu, and Wen Gao

School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, P.R. China
{ljwang,yhx,shaohl}@vilab.hit.edu.cn, wgao@ict.ac.cn

Abstract. The paper presents a novel video watermarking scheme called Mapping Energy video Watermarking (MEW) for copyright protection. The watermark is embedded in the compressed domain and extracted directly from the bitstream without original video. In the proposed scheme, we select a part of Integer DCT quantized coefficients to construct the embedded space and embed the watermark into it by using Energy Mapping Function (EMF). During the process of embedding, resynchronization strategy and watermarking coding are used to guarantee the robustness of watermarking. The experimental results indicate that the scheme has strong robustness to the attacks such as re-encoding, frame dropping, frame rate changing and sharpening. The influence of the coding efficiency is almost unnoticeable. Besides, high watermark payload and low time complexity are advantages of the scheme.

1 Introduction

With the development of computer network, copyright protection of digital products like video and audio has become a hot research topic. Video watermarking is an efficient technology to protect the copyrights of digital video. Many watermarking schemes based on spatial domain [1–2] and frequency domain [3–4] have been developed and can be used both in image and video. Compared with image watermark, video watermarking schemes should have their own characteristics such as the robustness to compression and the frame operations (dropping, re-encoding, shifting etc). Video is always disseminated as a compressed format, so a video watermarking scheme performed in the compressed domain is necessary. Hartung [5] proposed a typical scheme in the compressed domain. They arranged a spread-spectrum watermark with the same size as one frame and divided it into 8*8 blocks. Each block is transformed by DCT and added into the corresponding video DCT block. Cross [6] proposed the watermark was embedded in VLC of I-frame with the proper payload. Besides, the watermarking schemes performed in the compressed domain include those embedded in the residual of motion vectors [7–8] and in the facial parameters of MPEG-4 bitsteam [9]. These video watermarking schemes can gain low computational complexity or improve the robustness to the compression, but they are not resistant to such attacks as re-encoding and frame dropping. Lagendijk [10]

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 433–440, 2004.

has developed an algorithm called Extended Differential Energy Watermarking (XDEW), in which the watermark was embedded in both I-frames and P-frames. The algorithm is performed in the low bit-rate environment and has good performance on the robustness to re-encoding. But it is complicated from a computational standpoint. This paper proposes a novel video watermarking called adaptive Mapping Energy Watermarking (MEW) performed in the standard H.264 (JM 6.2) bitstream. We define the media data space is constructed by the quantized Integer DCT coefficients and the embedded space is a part of it. Every coefficient in the embedded space corresponds to one watermark bit and the value of the coefficient represents the energy of the watermark bit. This character of watermarking scheme gains it good performance on coding efficiency and watermark payload. The rest of this paper is organized as follows. In Section 2, MEW algorithm is explained in detail. In Section 3, the robustness of watermarking is analyzed and the experimental results are presented. Finally we present the conclusion of our experiments in Section 5.

2 MEW Algorithm

Video is a kind of media with a large data space and exits as a compressed format, so a video watermarking scheme in the compression domain is necessary. Figure 1 shows the diagram of our watermarking system based on the video encoder. From Figure 1, watermark and raw video are input as the original information and the watermarked compressed video and key (the quantized step is used to generate the key which guarantees the robustness to re-encoding at different bit rate) are output after embedding process; original watermark, watermarked video and the key are required during extracting process and the extracted watermark and detection value are output after extracting process.

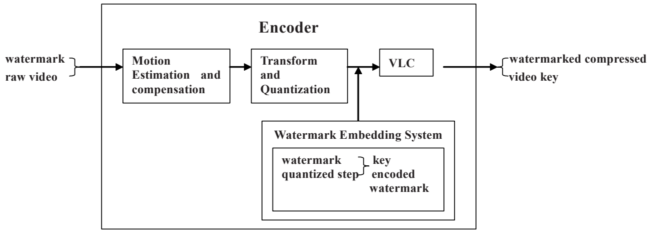


Fig. 1. The Diagram of Watermarking System

2.1 The Embedding Process

The embedding procedure is summarized as follows:

1. For a luma block, MInteger DCT quantized AC coefficients before a cut off point (cut_off) in zig_zag scanning order in a fixed scope ($a \leq abs(x) \leq b$) to construct the embedded space. If there are not as many as M coefficients that can be selected, this block would not be embedded.

2. The watermark can be any useful information, and in our experiments it is a random sequence of 1s and –1s generated by a key. Corresponding to Mselected coefficients, the watermark sequence is divided into several Groups of Watermark bits (GOW) with the size of M. Every selected coefficient corresponds to one watermark bit and we use Energy Mapping Function (formula 1) to embed the watermark.

$$f(x) = \begin{cases} x & a \leq abs(x) \leq c \quad ,w = 1 \\ -c & c+1 \leq x \leq b \quad ,w = 1 \\ c & -b \leq x \leq -(c+1) \;,w = 1 \\ x & c+1 \leq abs(x) \leq b \;,w = -1 \\ c & -c \leq x \leq -a \quad ,w = -1 \\ c & c \leq x \leq b \quad ,w = -1 \end{cases} \tag{1}$$

Where x is the value of one coefficient, w is one watermark bit, a and b are the bounds of embedded space and c is the partition point (in our experiments, a, b and c are 2,3 and 2).

3. The same GOW is embedded in the 4 luma blocks of one MB to improve the robustness of watermarking if the blocks can be embedded.

2.2 The Extracting Process

In contrast with the embedding process, we use the following procedure to extract watermark bits.

1. In one luma block, M coefficients ($a \leq abs(x) \leq b$) before a cut off point are selected. If there are not as many as M coefficients, the block isn't watermarked.

2. The energy of one watermark bit is calculated by adding all the corresponding de-quantized values that represent it as the formula 2 and 3.

3. The watermark bit can be obtained according to the sign of the energy as formula(4).

$$E_i = \sum (Esign(abs(de\_quant(x)))) \tag{2}$$

$$Esign = \begin{cases} -, & c+1 \leq abs(x) \leq b \\ +, & a \leq abs(x) \leq c \end{cases} \tag{3}$$

$$Esign = \begin{cases} 1, & E_k > 0 \\ -1, & E_k \leq 0 \end{cases} \tag{4}$$

In the formula 2, $de\_quant(x)$ represents the de-quantized value of coefficient x, $Esign(x)$ represents the sign of x according to its absolute value, E_k represents the energy value corresponding to the Kth watermark bit and $W_k$represents the Kth watermark bit.

As the formulas show, the embedding energy is enlarged to guarantee extracting watermark bits exactly because the sign of energy will retain unmodified even though some of coefficients value exceed its original scope during re-encoding.

3 Experimental Results

We test the performance of MEW in terms of watermark capacity, robustness and visual quality. We use H.264 bitstream coded at 30 fps, and the spatial resolution of the video sequences in our experiments is 176 × 144 pixels (QCIF). The watermark used in the experiments is a random sequence of 1s or –1s generated by a key.

3.1 Watermark Payload and Visual Quality

We compare the video quality with and without watermarking as Table 1. 1. Asnry, Asnru and Asnrv are the average of PSNR of three chroma components, which weigh the visual quality of the compressed video. 2. Embedded_bits are the amount of watermark bits embedded in the video, which weighs the watermark payload. 3. BIR is Bit Increased Rate that weighs the increased amount after watermarked.

Table 1. The experiments results of 4 standard test sequences 300 $frames$, 30fps, $cut\_off = 35, M = 8$

sequence parameters	Foreman_qcif		News_qcif		Silent_qcif		Akiyo_cif	
	original	Wm(1:1)	Original	Wm(1:1)	Original	Wm(1:1)	Original	Wm(1:1)
Asnry(db)	38.79	38.77	40.19	40.14	39.30	39.24	41.72	41.71
Asnru(db)	42.23	42.23	42.99	42.97	42.15	42.13	43.54	43.51
Asnrv(db)	43.95	43.94	43.62	43.63	42.97	42.93	44.56	44.58
Totalbits(bits)	3261856	3277552	1751176	1764440	1833064	1841386	726968	736928
Bit rate(kbps)	328.37	329.95	176.29	177.63	184.54	185.37	73.18	74.19
Total time(s)	218.408	227.689	193.778	209.255	208.983	207.750	191.128	198.176
Embedded_bits		2928		3176		2296		1736
BIR (%)		0.48116		0.76011		4.476		1.380
WBR(bps)		294.76		319.73		231.1		174.77

$$BIR = \frac{watermarked\_BR - original\_BR}{original\_BR} \times 100\% \tag{5}$$

Where $watermarked_BR$denotes the Bit Rate with watermark and $original_BR$denotes the Bit Rate without watermark. This parameter weighs the effect of compression efficiency caused by watermarking.

4.WBR denotes the Watermark Bit Rate, which weighs the watermark payload.

$$WBR = \frac{watermarked\_BR \times Embedded\_Bits}{total\_bits} \times 100\% \tag{6}$$

Where $total\_bits$ is the bits of watermarked video.

From Table 1, the coding efficiency of H.264 is very high and the degradation caused by watermarking is almost unnoticeable (the modification of PSNR and BIR are very small compared with encoding without watermarking). BIR is limited in 0.05 percent and WBR can get the largest payload of 300 bps (news, BR=177.63kbps). WBR can be large if the parameters are set properly.

Compared with XDEW algorithm, MEW has better visual quality and higher watermark payload. The following figures show PSNR and BIR curves. In fact, there is no law from the curves and the effect of PSNR and BIR is different with different video. But as a whole, PSNR is high and BIR is low.

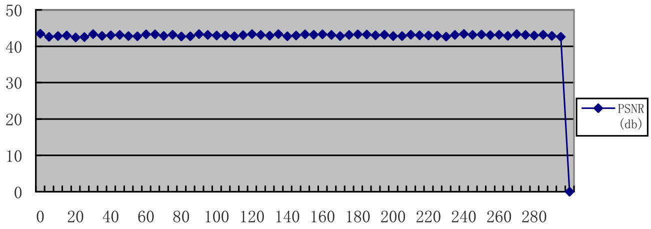


Fig. 2. PSNR of watermarked "Claire.yuv" video

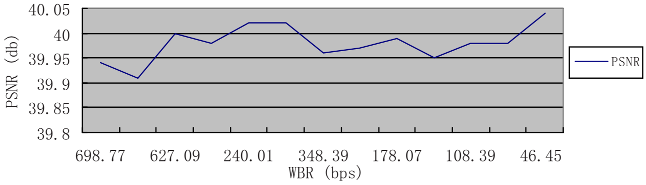


Fig. 3. PSNR at different WBR *News.yuv*

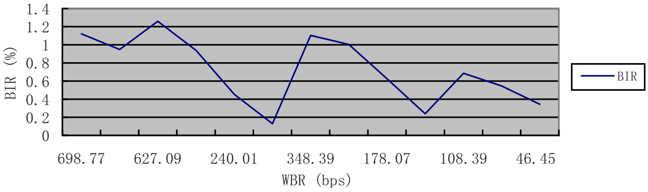


Fig. 4. BIR at different WBR *foreman.yuv*

3.2 The Robustness of Watermarking

As the experimental results prove, some non-watermarked MBs are included and some embedded MBs are discarded during embedding when the watermarked video is attacked by sharpening, frame dropping, re-encoding etc. When this scenario occurs, the synchronization between embedding and extracting is destroyed. We use re-synchronization strategy and watermark coding to guarantee the synchronization.

3.2.1 Re-synchronization Strategy

Re-synchronization strategy includes re-embedding and optimal retrieving strategies.

1. Re-embedding strategy:

As the description of embedding procedure, we embed a GOW in a luma block. For the continuous P MBs, we embed the same GOW in every luma block of them and these P MBs construct a Group of Embedded MBs (GEMB). When one MB is destroyed, GOW still can be extracted from other MBs in the same GEMB.

2. Optimal retrieving strategy:

We define a parameter as Dependable GOW (Dep_GOW) initiated by the correct GOW. For every Current extracted GOW ($Curr\_GOW$), the Similar Rate between Dep_GOW and $Curr\_GOW$ is calculated as formula 7. If SR is no less than T, the current MB with the extracted GOW belongs to the current GEMB and Dep_GOW is updated as formula 8; otherwise, the following Skipped Number (Max_SN) GOWs are extracted and compared each other, then the most similar two of Max_SN GOWs are found and the corresponding MBs belong to the next GEMB.

$$SR = (Dissimilar\_num) + Len(wm) = \begin{cases} \geq T, & , Same \\ < T, & , Different \end{cases} \tag{7}$$

Where $Dissimilar\_num$ denotes the number of different bits in the two GOWs and $len(wm)$denotes the length of one GOW (M). In our experiments, we set T as 0.75 to obtain good performance.

$$Dep\_GOW = Dep\_GOW + Curr\_GOW \tag{8}$$

Where Dep_GOWdenotes the dependable GOW, $Curr_GOW$denotes the current extracted GOW.

From the process described above, SR is the criterion of segmenting GEMBs.

3.2.2 Watermark Coding

When the watermarked video is attacked, the embedded space will be changed. We encode the watermark with Reed-Solomon coding due to its strong ability of correction. In our experiments, we set parameters of Reed-Solomon coding to correct at most 2 bits when 3 original bits are coded. Strong robustness can be obtained by sacrificing high watermark payload since the coded watermark bitstream is much larger than original watermark bitstream.

3.2.3 Experimental Results

In table 2, ABER denotes the Bit Error Rate after re-encoding which weighs the robustness of re-encoding (formula 9).

$$ABER = \frac{Error\_bits}{Embedded\_bits} \times 100\% \tag{9}$$

Where $Error\_bits$ denotes the number of error extracted watermark bits.

Table 2. The experimental results with R-S code for the 4 standard test video sequences at the fixed parameters $Max_SN = 2, = 4(176 * 144$ pixels, 30fps, 66 frames)

sequence parameters	Foreman_qcif		News_qcif		Silent_qcif		Akiyo_cif	
	original	Wm(1:1)	Original	Wm(1:1)	Original	Wm(1:1)	Original	Wm(1:1)
Asnry(db)	38.76	38.72	40.11	40.04	39.28	39.22	41.60	41.55
Asnru(db)	42.10	42.08	42.60	42.60	42.01	42.02	43.39	43.39
Asnrv(db)	43.85	43.86	43.41	43.41	42.93	42.91	44.46	44.44
Totalbits(bits)	596504	597880	356704	358336	400488	403536	174592	176336
Bit rate(kbps)	279.61	280.26	167.21	167.97	187.71	189.16	81.84	82.66
Total time(s)	181.312	185.579	164.456	179.369	165.766	175.925	154.373	176.854
Embedded_bits		12		15		12		2
BIR (%)		0.232		0.4545		0.7724		1.002
ABER (%)		46.67		51.136		35.577		62.5
ABER with RS coding (%)		0		0		0		0
WBR(bps)		5.625		7.0312		5.625		2.16

From Table 2, the effect of the performance of encoding efficiency caused by watermarking is very small compared with original encoding efficiency and the robustness to re-encoding with RS coding is very strong (ABER without RS coding is much larger than that with RS coding). From the point, Reed-Solomon code is very optimal to correct most error bits and it can gain the optimal performance if P is selected properly.

When the watermarked video is attacked, the synchronization between the embedding and extracting will be destroyed. Moreover, the effect of synchronization is different with different attacks.

Figure 5 shows BER curves at the attacks of re-encoding, frame rate changing, frame dropping and sharpening (original video is compressed at the quantized step of 27; re27 means the video is re-compressed at 27 and re10 means it is re-compressed at 10; frame dropping and sharpening mean the video is decompressed first, then an arbitrary frame is dropped and all the I frames are sharpened, finally re-compressed at 27).

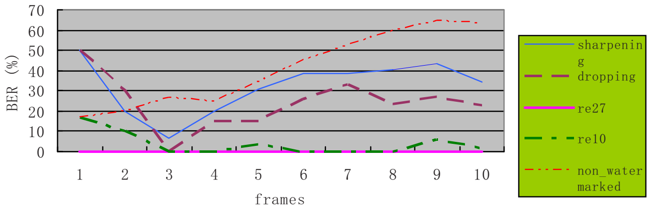


Fig. 5. BER at the attacks of sharpening, frame dropping and re-encoding *News.yuv*

Where non_watermark curve represents the BER of video without watermark. From the figure 5, BER is limited in 20 when the video is re-compressed at quantized step of 27 and 10. BER is oppositely large at first frames in dropping and sharpening curves, but it is decreased when I frames are increased.

4 Conclusion

This paper presents a novel blind video watermarking algorithm MEW. MEW algorithm includes re-embedding, optimal retrieving and watermark coding. These strategies are used to guarantee the encoding efficiency and the robustness of watermarking. As the experimental results prove, the performance of MEW is very excellent. MEW algorithm has such advantages as low time complexity, large watermark payload and strong robustness to re-encoding. We only use MEW to embed the watermark in I frames, in fact, P or B frames can also be used to embed more watermark bits or improve the robustness of watermarking. We use H.264 compression standard to test our algorithm and it can be extended to other standards as MPEG-2, MPEG-4 etc.

References

1. R.van Schyndel, A. Tirkel and C. Osborne: A digital watermark. IEEE ICIP, Vol.2, (1994) 86–90
2. B.C. Mobasseri, M. J. Sieffert and R. J. Simard: Content authentication and tamper detection in digital video. IEEE ICIP, Vol.1, (2000) 458–461
3. I.J.Cox, J.Killian, T.Leighton and T.Shamoon: Secure spread spectrum watermarking for multimedia. IEEE Trans. On Image Processing, Vol.6 (12), (1997) 1673–1687
4. Xiaoyun Wu, Wenwu Zhu, Zixiang Xiong and Ya-Qin Zhang: Object-based multiresolution watermarking of images and video. IEEE ISCS 2000, Vol.1, (2000) 212–215
5. F. Hartung and B. Girod: Digital watermarking of MPEG-2 coded video in the bitstream domain. ICASSP-97, Vol.4, (1997) 2621–2624
6. D. Cross and B. G. Mobasseri: watermarking for sef-authentication of compressed video. IEEE ICIP, Vol.2, (2002) 913–916
7. Zhuo zhao, Nenghai Yu and Xuelong Li: A novel video watermarking scheme in compressed domain based on fast motion estimation. IEEE ICCT, (2003)1878–1882
8. Yuanjun Dai, Lihe Zhang and Yixian Yang: A new method of MPEG video watermarking technology. IEEE ICCT, (2003)1845–1847
9. F. Hartung, P. Eisert and B. Girod: Digital watermarking of MPEG-4 facial animation parameters. Computers and Graphics, Vol. 22, (1998) 425–435
10. I. Setyawan and R. L. Lagendijk: Low bit rate video watermarking using temporally extended differential energy watermarking(DEW) algorithm. Proc. Security and Watermarking of Multimedia Contents III, Vol. 4314, (2001) 73–84

A Fragile Image Watermarking Based on Mass and Centroid

Hui Han, HongXun Yao, ShaoHui Liu, and Yan Liu

School of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, P.R. China
{hh,yhx,shaohl,lyan}@vilab.hit.edu.cn

Abstract. In this paper, a novel fragile watermarking is proposed. It is different from the common fragile watermarking, which uses the cryptographic digital signature of an image or image blocks as authentication information, because two new conceptions, mass and centroid in physics are introduced into our scheme and the fragile watermarking is designed based on them. The mass and the position of centroid are used as authentication information for the image block. Since two random sequences are applied to randomize the abscissa and ordinate of an image block's centroid, the algorithm can successfully avoid the drawback of cryptographic hash function, which suffers the famous birthday attack. When attackers tamper the image, the mass and the position of centroid are changed, with the changes of them, we can effectively locate the tamper done to the image. Experimental results show that this watermarking can detect and locate any change made to an image.

1 Introduction

With the rapid development of computer science and network technology, communication and exchange of multimedia information have stretched to an unprecedented depth and width. Multimedia information, however, often suffers from all kinds of accident or deliberate tampering, which makes people become suspicious about the integrity and authenticity of images, audio and video. If tamper touches upon state security, evidence provided in the court and historic documents, it may cause negative social influence and major political or economic loss. Therefore, how to verify the integrity and authenticity of digital media has become a severe problem. Fortunately, as a branch of multimedia watermarking technology,fragile watermarking provides a solution to this problem for us. Fragile watermarking is usually applied to integrity authentication and local tamper detection. Any manipulation done to multimedia will lead to incorrect watermark extraction resulting in verification failure.

Recently, many fragile watermarking techniques have been proposed for the purpose of integrity authentication and local tamper detection. The earliest fragile watermarking was based on least significant bit (LSB)[1][2]. Even if a single pixel is changed the watermark will be destroyed, thus the changed area can be located. However, the disadvantage of this method is that attackers can modify

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 441–448, 2004.

images by changing the bits on other bit planes while keep their LSB unchanged. Wong [3][4] propose a method which partitions the image into sub-blocks and then use a cryptographic hash function such as MD5, to compute the digital signature of each image block and the size of the whole image. With the signature as authentication information, the scheme embeds watermark into the LSB of every image block. The improved algorithms were proposed in [5][6][7].

In this paper, we propose a new fragile watermarking based on the mass and centroid of an image. The rest of this paper is structured as follows. Both concepts of mass and centroid are briefly introduced in Section 2. And then the embedding algorithm will be described in Section 3. In Section 4, the watermark extraction and verification will be presented. Finally, experimental results are given in Section 5 followed by the conclusions in Section 6.

2 Mass and Centroid of Image and Calculation

It is well known that every object in the universe has its mass and centroid. Mass reflects how much substance it contains and centroid shows the centralized position of its substance. Take a discrete substance system in a rectangular coordinate system for example, we suppose the mass of every point is $m_i(i = 1, 2 \cdots\cdots n)$, and the coordinate of every point is (x_i, y_i). Its total mass is $M = \sum\limits_{i=1}^{n} m_i$ and its centroid is (X, Y) where $X = \frac{1}{M}\sum\limits_{i=1}^{n} m_i \times x_i$, $Y = \frac{1}{M}\sum\limits_{i=1}^{n} m_i \times y_i$. One digital grayscale image is obtained by taking discrete samples of a continuous image. Correspondingly, a discrete grayscale image can be regarded as the discrete substance system in physics and the grayscale level of every pixel is the mass of this point. Thus, an image has its mass and centroid. Let I denote a grayscale image of size $m \times n$ and a rectangular coordinate system is established for I. The point (1,1) in the rectangular coordinate system corresponds to the pixel at left bottom corner and the point (m, n) is the pixel at the right up corner. The level distance between two pixels is the unit length of the rectangular coordinate system. In this way, all the pixels have their corresponding integer coordinates (see Fig.1). Let $f(x, y)$ represent the grayscale level of the pixel at point (x, y) in the rectangular coordinate system. The total mass is M and the centroid coordinate is (X, Y). The formulas are:

$$M = \sum_{x,y=1}^{m,n} f(x, y) \tag{1}$$

$$X = \frac{1}{M} \sum_{x,y=1}^{m,n} f(x, y) \times x \tag{2}$$

$$Y = \frac{1}{M} \sum_{x,y=1}^{m,n} f(x, y) \times y \tag{3}$$

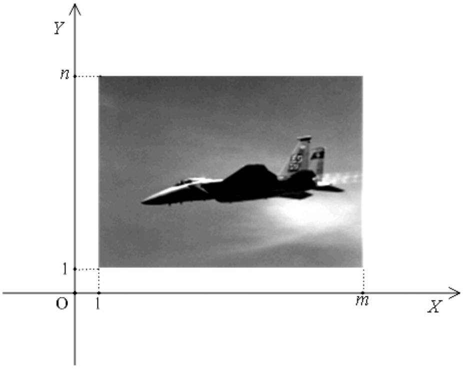


Fig. 1. Image in coordinate

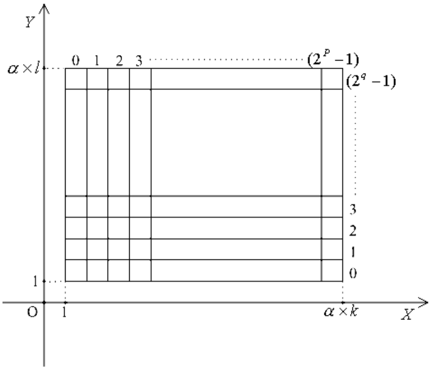


Fig. 2. Dividing the coordinate

Obviously, M in the formulas means the sum of all grayscale level in this area and (X, Y) represents the centralized position of grayscale.

3 Embedding Algorithm

Since an image has its mass and centroid, image blocks also have mass and centroid. The mass and centroid of image blocks are used as authentication information to verify every image block one by one. According to formulas (1), (2), (3), once attackers know the grayscale level of every pixel in an image block, it is very easy for them to calculate the mass and centroid. Hence if attackers use these formulas to attack the watermarking system, the system is vulnerable. For the sake of security, we will improve the centroid formula when we verify each image block.

Let A denote the original $m \times n$ grayscale image. Image A is partitioned into non-overlapping sub-blocks. These sub-blocks are arranged in raster-scan sequence with A_i denoting the i^{th} block. $(i = 1, 2, \cdots\cdots \left\lceil \dfrac{m \times n}{k \times l} \right\rceil)$. We generate two integer random matrixes RMX and RMY of size $m \times n$ with $K1$ and $K2$ as the private keys. The value of Elements in RMX and RMY is integer within $1 \sim \alpha$. Then they are partitioned into non-overlapping matrix of size $k \times l$ just like partitioning A. Let RMX_i and RMY_i denote the i^{th} matrix respectively. RMX_i and RMY_i will be used to improve the centroid calculation formulas of image blocks. We use a binary matrix W(its elements are 0 or 1) of size $m \times n$ as the watermark information and partition it into sub-matrix W_i of size $k \times l$ corresponding to the image sub-block A_i.

3.1 Improved Centroid Calculation Formulas

The derivation of the improved centroid calculation formulas for image blocks can be summarized as following three steps:

Step 1. We form the corresponding block $\overline{A_i}$ where each element in $\overline{A_i}$ equals the corresponding element in A_i except that the LSB is set to 1. And then the corresponding rectangular coordinate system is established for image block $\overline{A_i}$ just like establishing coordinate for images. That is, the point (1,1) corresponds to the pixel at the left bottom corner and the point (k, l) corresponds to the pixel at the right up corner of the image block. Let $f_i(x, y)$ stand for the grayscale level of the pixel whose coordinate is (x, y).

Step 2. We calculate the mass of image block $\overline{A_i}$, $M_i = \sum_{x,y=1}^{k,l} f_i(x, y)$. The maximum of pixel grayscale level is 255, so obviously $M_i \leqslant 255 \times k \times l$, M_i can be represented by $s = \lceil \log_2(255 \times k \times l + 1) \rceil$ bits.

Step 3. Let (X_i, Y_i) denote the centroid's coordinate of image block $\overline{A_i}$ and (RX_i, RY_i) denote the improved centroid's coordinate. Now, the improved centroid (RX_i, RY_i) of image block $\overline{A_i}$ can be represented as:

$$RX_i = \frac{1}{M_i} \sum_{x,y=1}^{k,l} f_i(x, y) \times RMX_i(x, y) \times x \tag{4}$$

$$RY_i = \frac{1}{M_i} \sum_{x,y=1}^{k,l} f_i(x, y) \times RMY_i(x, y) \times y \tag{5}$$

3.2 Locating The Improved Centroid

It is well known that the centroid of a solid convex object should be inside this object. Hence the abscissa and ordinate of centroid (X_i, Y_i) of image block $\overline{A_i}$ should fall into the interval $[1, k]$ and $[1, l]$ respectively. From the formulas (4) and (5), it is easy to see that the abscissa of the improved centroid (RX_i, RY_i) should belong to interval $[1, \alpha \times k]$ and the ordinate should belong to interval $[1, \alpha \times l]$. The process of locating the centroid is described as follows:

Definition 1. The minimum variation rate of the improved centroid's position. Because the variation of substance distribution will lead to the change of an object's centroid, similarly the change of pixel grayscale will result in the change of the improved centroid's position of image blocks. For example, the variation of grayscale level of the pixel whose coordinate is (c, d) in the i^{th} image block is Δ, its corresponding random variables are $RMX_i(c, d)$ and $RMY_i(c, d)$. The mass of the image block is M_i. Then the variation of the improved centroid's abscissa caused by the change of this point's grayscale level is $\Delta \times c \times RMX_i(c, d)/M_i$ and the variation of its ordinate is $\Delta \times d \times RMY_i(c, d)/M_i$. It is evident that if Δ, c or d, $RMX_i(c, d)$ or $RMY_i(c, d)$ equals 1 in the above formulas and its mass M_i is $255 \times k \times l$, the variation of its abscissa and ordinate is the smallest: $r_{\min} = \dfrac{1 \times 1 \times 1}{255 \times k \times l}$. We define $r_{\min} = \dfrac{1}{255 \times k \times l}$ as the minimum variation rate of an image block's centroid.

If the minimum variation rate of the improved centroid $r_{\min}$ can be detected, certainly other variation can be detected. In order to detect the variation of the improved centroid, the abscissa and ordinate interval of the improved centroid are divided. At first, we divide the abscissa interval of the improved centroid $[1, \alpha \times k]$ into 2^p intervals equally and index each interval from 0 to $(2^p - 1)$ (See Fig. 2). Now every interval can be represented by p bits, where $p = \left\lceil \log_2(\frac{\alpha \times k - 1}{r_{\min}}) \right\rceil$. Obviously, the width of every interval $(\alpha \times k - 1)/2^p$ is less than or equals $r_{\min}$. As for the ordinate, we can find an integer $q = \left\lceil \log_2(\frac{\alpha \times l - 1}{r_{\min}}) \right\rceil$ and divide the ordinate interval of the improve centroid $[1, \alpha \times l]$ into 2^q intervals equally. It is apparent that the width of every interval $(\alpha \times l - 1)/2^q$ is less than or equals $r_{\min}$. We index these intervals from 0 to (2^q-1)(see Fig. 2) and use q bits to represent each interval. The abscissa and ordinate of the improved centroid must fall in a certain interval or on the boundary of intervals we have divided. When it falls on the boundary, it is considered in the previous interval. Because the width of every interval is less than or equals $r_{\min}$, when the grayscale level of an image block changes, its improved centroid's abscissa and ordinate must fall in other intervals or on other boundaries. Therefore, we can make sure that the improved centroid's position has been changed.

Let IDX_i denote the index of the interval which the improved centroid abscissa fall in and IDY_i denote the index of the interval which the improved centroid ordinate fall in, then we can get:

$$IDX_i = \left\lfloor \frac{RX_i}{(\alpha \times k - 1)/2^p} \right\rfloor - 1 \tag{6}$$

$$IDY_i = \left\lfloor \frac{RY_i}{(\alpha \times l - 1)/2^q} \right\rfloor - 1 \tag{7}$$

3.3 Watermark Insertion

This process involves generating embedding information and inserting it into the LSB of image block $\overline{A_i}$. First, we represent M_i, IDX_i, IDY_i using s, p, q bits and arrange them into a $k \times l$ binary matrix P_i. (If $s + p + q < k \times l$, use 0 to fill the vacancy. If $s + p + q > k \times l$, the bits that are more than $k \times l$ bits will be discarded.) We combine P_i with W_i using an exclusive or function. That is, we compute $\bar{W}_i = P_i \oplus W_i$ where $\oplus$ denotes the element-wise exclusive OR operation between the two blocks. Then we use private key K to encrypt $\overline{W}$, $C_i = E_k(\bar{W}_i)$ and insert C_i into the LSB of image block $\overline{A_i}$.

4 Watermark Extraction and Verification

It has been known that the change of pixel grayscale level will lead to the change of improved centroid's position and mass. When we verify image blocks, by only

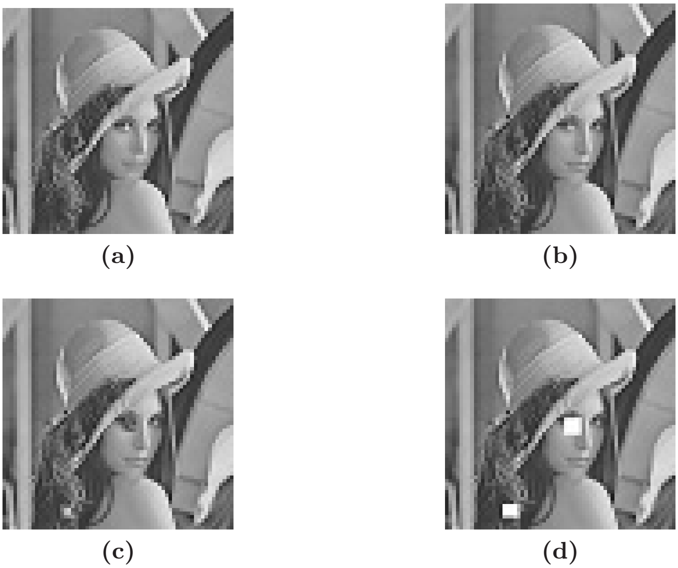

(a) (b) (c) (d)

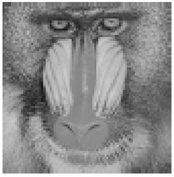
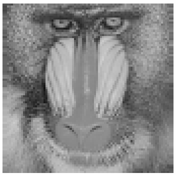
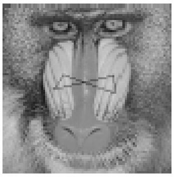
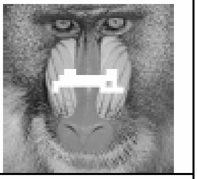
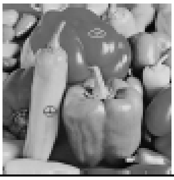
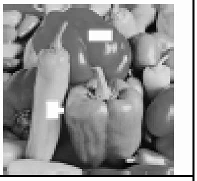
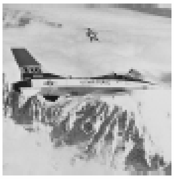
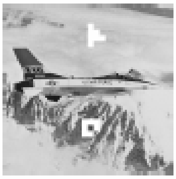
Fig. 3. (a) Original image (b) Watermarked image (c) Tampered watermarked image (d) Detection result

detecting whether the position of the improved centroid or mass is changed or not, we can find out whether pixel grayscale level changes or not. We partition the image into blocks just like embedding algorithm and split the verification image block V_i into two parts Z_i and Q_i. Z_i is the LSB part of the image block. For Q_i, each element in Q_i equals the corresponding element in V_i except that the LSB is set to 1. After the same processing in embedding algorithm exerted on Q_i, we represent M_i', IDX_i' and IDY' using s, p, q bits and arrange them into $k \times l$ binary matrix P_i'. M_i' is the mass. IDX_i' and IDY' are the index of the interval which the improved centroid's abscissa and ordinate fall in respectively. Then we use private key K to decipher Z_i, $U_i = D_k(Z_i)$ and compute $E_i = U_i \oplus P_i'$ using an element-wise exclusive or procedure. It is obvious that if E_i is the same with its corresponding embedding watermark block W_i, there is no variation happened to the mass and position of the improved centroid and image block V_i pass the verification. Conversely, this image block does not pass the verification.

5 Experimental Results

To demonstrate the feasibility of our scheme, a test has been done. The test we do is that random tampering the image. The famous Lena image of size 512×512 is used as the test image. We apply the method described above to partition the image into image blocks of size 8×8 equally and generate two random matrixes RMX and RMY with 512×512 elements by the private keys. The value of elements in them is integer within $1 \sim 128$. As for every image block, we use

Table 1. Experimental results

Original image	Watermarked image	Tampered watermarked image	Detection result
			
			
			

14 bits to represent the mass of an image block and 24 bits to represent the abscissa and ordinate position information of the improved centroid respectively to form authentication information. Private key symmetric encryption function has been applied to encrypt the bits series, which generated by the combination of the watermark and the authentication information. (Certainly, we can employ some more complicated encryption algorithms to encrypt the bits series to make the bits series more secure.) The following images are experimental results. Figure 3(a) is the original 512×512 grayscale Lena image. The watermarked image is given in Fig. 3(b). We can see that the watermarked image has good quality with a peak-signal-noise rate (PSNR) of 51.0db.The tampered watermarked image, where we tamper the hair of the girl on the back of her shoulder and the eye of her is shown in Fig. 3(c). Figure 3(d) is the detection result of tampered area. It indicates that our scheme can successfully detect the tamper. (The white area in the detection result means the places where the image has been tampered.) We also use the same way to test other images, experimental results are shown in Table 1. From the results, we can draw the conclusions that our method can effectively detect the tampering for all kinds of images.

6 Conclusion

In this paper, we present a novel watermarking scheme to authentication for images. The image authentication method is based on the mass and centroid of images. The algorithm can effectually verify every image block by detecting the variation of the mass and centroid, thus the tampered area can be located. Experimental results show this kind of fragile watermark can detect effectively any tamper done to the image and locate the tamper without the original image.

References

1. Gary L. Friedman: The Trustworthy Digital Camera: Restoring Credibility to the Photographic image. IEEE Transactions on Consumer Electronics, Vol.39, No.4. (1993) 905–910
2. R. G. van Schyndel, A. Z. Tirkel, C. F. Osborne: A Digital Watermark. IEEE International Conference on Image Processing, Vol.2. Austin, Texas, USA (1994) 86–90
3. P. W. Wong: A Public Key Watermark for Image Verification and Authentication. IEEE International Conference on Image Processing, Vol.1. Chicago, Illinois, USA (1998) 455–459
4. P. W. Wong: A Watermark for Image Integrity and Ownership Verification. IS&T PIC Conference. Portland, Oregon, USA (1998) 374–379
5. Paulo S. L. M. Barreto, Hae Yong Kim, Vincent Rijmen: Toward A Secure Public-key Blockwise Fragile Authentication Watermarking. IEEE International Conference on Image Processing. Thessaloniki, Greece (2001) 494–497
6. LIU Feilong, WANG Yangsheng: An Improved Block Dependent Fragile Image Watermark. IEEE International Conference on Multimedia & Expo, Vol.2. Baltimore, Maryland, USA (2003) 501–504
7. Mehmet Utku Celik, Gaurv Sharma, Eli Saber, Ahmet Murat Tekalp:Hierarchical Watermarking for Secure Image Authentication with Localization. IEEE Transactions on Image Processing, Vol. 11, No. 6. (2002) 585–595

MPEG-21 DIA Testbed for Stereoscopic Adaptation of Digital Items

Hyunsik Sohn, Haksoo Kim, and Manbae Kim

Kangwon National University
Department of Computer, Information, and Telecommunication
192-1 Hoja2-dong, Chunchon 200-701, Republic of Korea
manbae@kangwon.ac.kr

Abstract. MPEG-21 Digital Item Adaptation (DIA) provides digital items being adapted according to user preferences and terminal capabilities. This paper considers the implementation of stereoscopic adaptation in the DIA Testbed being composed of a DIA server, a client, and a network interace module. User preferences considered for the stereoscopic adaptation are the types of stereoscopic parallax, the range of 3-D depth, and the interval of a previous frame. Such descriptors are sent to the DIA server in the form of XML. Then, the server adapts the descriptors as well as resources and transmits them to the client. At the server side, MPEG-1 video is converted into stereoscopic MPEG-4 video. Upon receiving the streamed video in RTP, the client displays the stereoscopic video. RTP/RTSP and TCP/IP protocols are used to deliver various types of data between the client and server.

1 Introduction

The vision for MPEG-21 is to define a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities [1]. Digital Item is a structured digital object including resource and and descriptor, which is a fundamental unit of distribution and transaction within MPEG-21 multimedia framework. Digital Item Adaptation (DIA) is one of main MPEG-21 parts. The goal of the DIA is to achieve interoperable transparent access to multimedia contents by shielding users from network and terminal installation, management and implementation issues. This will enable the provision of network and terminal resources on demand to form user communities where multimedia content can be created and shared.

There are a variety of digital items and their associated adaptation methods. For instance, the adaptation of image, video, audio, graphics, etc is introduced in the standard. Among various digital items, this paper considers stereoscopic adaptation or conversion, in which 2-D video is adapted or converted into stereoscopic video (2D-to-3D stereoscopic conversion). For its practical application, we propose MPEG-21 DIA Testbed that implements such adaptation over a network. Our proposed DIA Testbed is composed of a DIA server, a client, and a

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 449–456, 2004.

network interface module (NIM) where DI resources are adapted according to user-preference descriptors defined by MPEG-21 DIA [3]. The descriptors that manage the stereoscopic adaptation are *the types of parallax*, *the range of depth*, and *the interval of a previous frame*. Client's descriptors are delivered to a DIA server across a network. Then the DIA server transports adapted DI to the client using RTP (Real-time Transport Protocol)/RTSP (Real-time Transport Streaming Protocol) for real-time streaming [7,8].

This paper is organized as follows. The following section describes the overall architecture of DIA Tested. Section 3 presents s stereoscopic conversion method and some experimental results. Section 4 describes the structure of a network interface module. The implementation of the Testbed is presented in Section 5 followed by the conclusion of Section 6.

2 Overview of DIA Testbed

Figure 1 shows how the stereoscopic adaptation is carried out as defined in DIA standard. The digital item is separated into D (descriptor) and R (resource) by DEMUXER. In the Resource Adaptation, the 2-D video resource, R is converted into stereoscopic video, R' by three stereoscopic descriptors such as *ParallaxType*, *DepthRange*, and *MaxDelayedFrame* in the *Descriptor Adaptation* block. Similar to R, D is also modified to a new D'. The new descriptor D' and resource R' form a stereoscopic digital item that would be delivered to a client over a network.

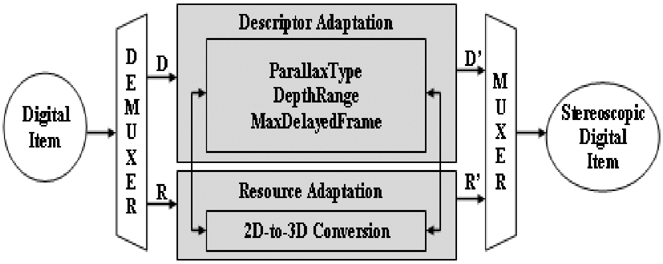


Fig. 1. The stereoscpic adaptation of DIA

The overall architecture of our proposed DIA Testbed whose main parts are a DIA server, a client, and a network interface module (NIM) is shown in Fig. 2. At the side of the client, main modules are a *User Preference* that chooses user preferences, a *Content Digital Item* (CDI) *Description Generator* that produces CDI.xml, a *conteXt Digital Item* (XDI) *Description Generator* that produces XDI.xml, and *Digital Item Player* (DIP) that displays a stereoscopic DI. Three descriptors stored in the XDI.xml are *ParallaxType, DepthRange*, and *MaxDelayedFrame*. Besides, the resource is stored in CDI.xml. The client requests an adapted DI using DIA Negotiation Message (DIANM) of CDI.xml and

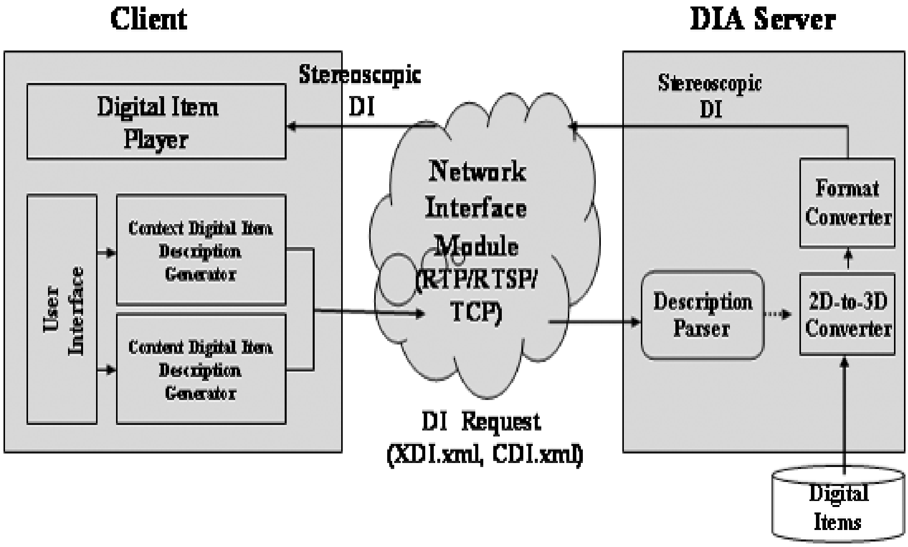

Client
DIA Server
Stereoscopic DI
Digital Item Player
Network Interface Module (RTP/RTSP/ TCP)
User Interface
Context Digital Item Description Generator
Content Digital Item Description Generator
Stereoscopic DI
Format Converter
Description Parser
2D-to-3D Converter
DI Request (XDI.xml, CDI.xml)
Digital Items

Fig. 2. DIA Testbed is composed of a client, a DIA server, and a network interface module (NIM)

XDI.xml [6]. The XML files are transmitted to the DIA server through NIM. The DIA server consists of a description parser that parses the XML files, a 2D-to-3D converter, and a format converter for MPEG-1 to MPEG-4 transcoding. RTP and RTSP are used for streaming and transmission controls of MPEG-4 resource, respectively. As well, the XDI.xml containing the descriptors is delivered in TCP/IP.

3 Stereoscopic Conversion Method

This section presents a methodology underlying a stereoscopic conversion by its associated stereoscopic descriptors as well as some experimental results. There are reported various conversion schemes [4,5]. One of stereoscopic conversion schemes is to make use of a delayed (previous) image. Based on this, we will describe how 2-D video is converted into a stereoscopic video under the descriptors.

Suppose that an image sequence is $\{\cdots, I_{K-3}, I_{K-2}, I_{K-1}, I_K, \cdots\}$ and I_K is a current frame. One of previous frames, I_{K-i} $(i \geq 1)$ is chosen. Then, a stereoscopic image consists of I_K and I_{K-i} . If the current and previous images are appropriately presented to both human eyes as in Table 1, then the user feels the 3-D stereoscopic perception [9]. *MaxDelayedFrame* determines the amount of i value. Thus, the larger it is, the more depth the user feels. Figure 3 shows four stereoscopic images that were converted from an MPEG-1 2-D *Fun* sequence, where MaxDelayedFrame is varied in [1,4]. The top field of a left image and the bottom field of a right image are interlaced. It is observed that the

Table 1. The selection of left and right images according to camera and object motions

Camera Motion	Object Motion	Left Image	Rigth Image
Right	None	Previous	Current
Left	None	Current	Previous
None	Right	Current	Previous
None	Left	Previous	Current

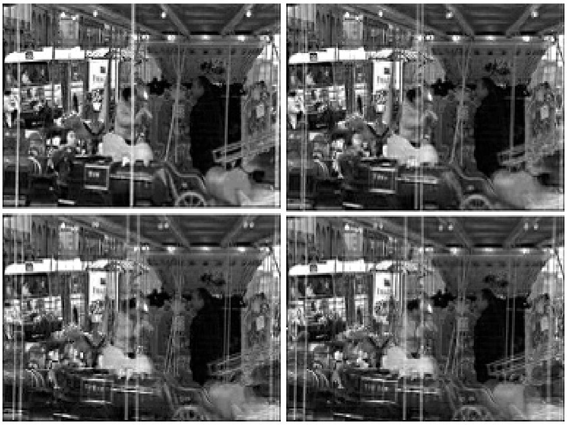


Fig. 3. Stereoscopic images with varying MaxDelayedFrame in [1,4]

disparity between the two images increases being proportional to the value of MaxDelayedFrame.

ParallaxType represents the types of parallax being composed of positive parallax and negative parallax [10]. This descriptor would control the positive and negative parallaxes in the stereoscopic adaptation of 2-D video. For a simple example, we use an MPEG-1 2-D *Flower and Garden* sequence. Figure 4 shows the interlaced stereoscopic images with positive parallax or negative parallax that could be generated by switching left and right images. In (a), a current image and a previous image are displayed to left and right eyes, respectively, where 3-D depth is perceived in the negative parallax. By switching the two images like in (b), the scene is perceived in the positive parallax.

DepthRange indicates the range of 3-D depth perceived by the user and is defined as the distance between a monitor screen and a 3-D location of an object. The distance could be normalized at [0,1]. The DepthRange applies identically to the positive and negative parallaxes. For the positive parallax, the range of depth increases by shifting a right image to the left direction with a left imaged fixed. On the contrary, shifting it to the right direction decreases the range of depth. Figure 5 shows the interlaced stereoscopic images with varying DepthRange. As it increases, a greater disparity between the two images is observed, thus producing more 3-D depth range.

Fig. 4. Two stereoscopic images with different parallax types. (Left) negative parallax and (Right) positive parallax

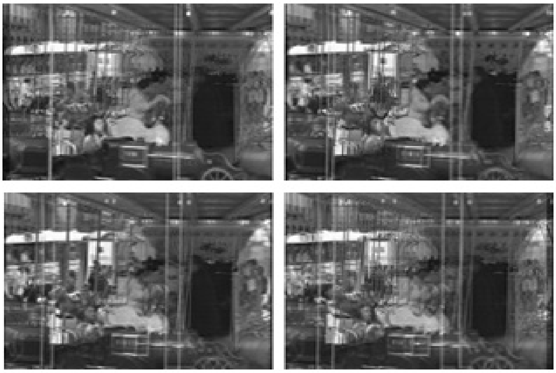


Fig. 5. Stereoscopic images with varying DepthRange of 0.05, 0.35, 0.65, and 0.95

4 Network Interface Module

Protocols directly related to DIA Testbed can be classified as network layer protocol, transport protocol, and session control protocol. Figure 6 shows the protocol stacks illustrating their relationship. At the server side, the resource of A/V DI is packetized at the RTP layer. The RTP packets are streamed to the UDP layer and the IP layer. The final IP packets are transported over the Internet. For the stream control, the client's control signals are packetized and streamed to the RTSP, TCP and IP layers. CDI.xml and XDI.xml are transported to the client using the TCP layer and the IP layer.

In the DIA Testbed, RTP and RTSP are used for the real-time transmission of stereoscopic digital items. RTP provides the following functions in support of media streaming: timestamping, sequence numbering, payload type identification, source identification, and so forth. RTSP is a session control protocol for streaming media over the Internet [8]. One of the main functions of RTSP is to support VCR-like control operations such stop, pause/resume, fast forward, and fast backward. In addition, RTSP also provides means for choosing delivery

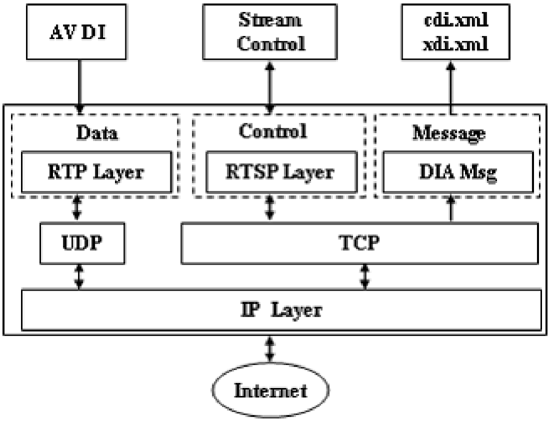


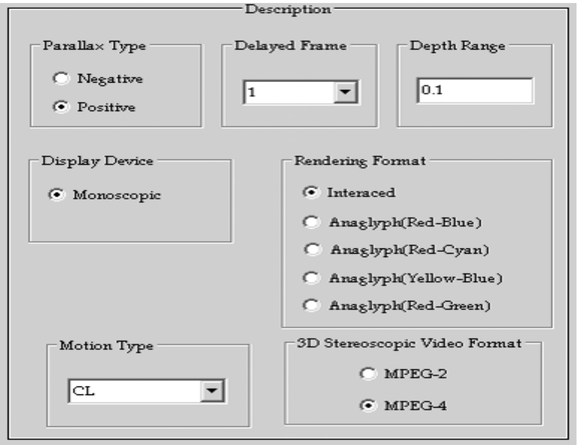
Fig. 6. Protocol stacks of DIA Testbed

channels (e.g., UDP, multicast UDP, or TCP) and delivery mechanism based upon RTP. In RTSP, each presentation and media stream is identified by an RTSP universal resource locator (URL). The following four RTSP messages are used: DESCRIBE, SETUP, PLAY, and TEARDOWN [8].

1. DESCRIBE: On receiving, the DIA server establishes a connection with a client, and examines whether a digital item is present at the server at RTSP URL. SDP parser extracts SDP and other information is sent to the client.
2. SETUP: DIA server sets port numbers of the server and client.
3. PLAY: The digital item is packetized and transmitted to the port number set by SETUP.
4. TEARDOWN: DIA server terminates the connection with the client and releases any allocated resources. Besides, the server returns to a wait state for the future connection with the client

5 Experiments

The DIA server has an MPEG-1 decoder, a 2D-to-3D converter, a Negotiation Message (NM) parser, and an MPEG-4IP processing unit. NIM has RTP/RTSP /UDP and TCP/IP modules. The client has a user preference window, a NM parser, and a Digital Item Player (DIP). MPEG-1 encoded sequences are stored in the DIA server. If the client transmits NM over the Internet, then the DIA server converts an MPEG-1 video into a stereoscopic MPEG-4 video and subsequently MPEG-4IP for video streaming. Besides, NM is also modified and transmitted to the client at the TCP/IP layer. Finally, the client parses the received NM and DIP playbacks the stereoscopic video. Most of functions are written in C/C++. MPEG-4 RTP packetization modules are based upon Darwin Streaming Server. The client modules are written based on MPEG-4IP wmplayer. In the



Description

Parallax Type
○ Negative
● Positive

Delayed Frame
1

Depth Range
0.1

Display Device
● Monoscopic

Rendering Format
● Interaced
○ Anaglyph(Red-Blue)
○ Anaglyph(Red-Cyan)
○ Anaglyph(Yellow-Blue)
○ Anaglyph(Red-Green)

Motion Type
CL

3D Stereoscopic Video Format
○ MPEG-2
● MPEG-4

Fig. 7. The client DIANM window

Fig. 8. DIP of a client showing stereoscopic video streamed over an Internet

design of the software, the following four classes are defined: CRTSPResponse, CRTSPmsg, CMP4RTP and CRTPServer. CRTSPResponse class manages the response on RTSP messages. CRTSPmsg class parses RTSP messages. Then, CRTSPResponse class sends its response to the client. CMP4RTP class analyzes MP4 file and produces RTP payload. CRTPServer class is a main one that establishes a connection with the client.

Figure 7 shows the client's user preference window setting DIA NM that would be delivered to the DIA server. 2D-to-3D stereoscopic conversion menu has ParallaxType, MaxDelayedFrame and DepthRange which are represented by Parallax Type, Delayed Frame and Depth Range, respectively. Additionally, Rendering Format indicates five formats of stereoscopic video which are interlaced and four other anaglyph modes. When setting MPEG-4 button active, the stereoscopic video is encoded in MPEG-4IP at the side of DIA server. CDI.xml and XDI.xml are transmitted to the DIA server. Then, the DIA server parses the two XML files and adapts the MPEG-1 video to stereoscopic MPEG-4 video according to the descriptors of XDI.xml. Figure 8 shows the DIP window of MPEG-4 video adapted according to user preferences.

6 Conclusion

In this paper, we presented the MPEG-21 DIA Testbed focusing on the stereoscopic adaptation of 2-D video in the network environments. For the implementation, we developed a testebed being composed of a DIA server, a client, and an NIM. The DIA server consists of a description parser, a 2D-to-3D converter, and a format converter and was designed to convert 2-D video to 3-D stereoscopic video according to descriptors sent by the client. At the client side, main modules are a user preference, a CDI generator, an XDI generator, and a Digital Item Player. For the efficient transmission of DIs, RTP/RTPS/UDP and TCP/IP are used because various types of data require appropriate network protocols. The adapted stereoscopic video is packetized at RTP layer and transmitted to the client in the Internet. It was demonstrated that adapted video are properly streamed to the client.

Acknowledgments. This work was supported by the Ministry of Information and Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.

References

1. "MPEG-21 Overview V.4", ISO/IEC JTC1/SC29/WG11 N4801, May 2002
2. ISO/IEC Draft of Technical Reports 21000-1, "Part1: Vision, Technologies and Strategy", MPEG/N4333, July 2001
3. MPEG-21 Digital Item Adaptation CD 21000-7 ISO/IEC JTC1/SC29/WG11/ N5612 March 2003, Pattaya, Thailand
4. T. Okino et al., "New television with 2-D/3-D image conversion technologies," SPIE Vol, 2653, Photonic West, 1995
5. Man Bae Kim, Jeho Nam, Woonhak Baek, Jungwha Son, Jinwoo Hong, "MPEG-21 DIA Description for 3-D Stereoscopic Video Conversion", IWAIT January 21-22, 2003, Nagasaki, Japan
6. J. Sohn et al., "The implementation of MPEG-21 Testbed in WEB environments", J. of Info. and Telecom Equipment, Vol. 1, No. 2, 2002
7. IETF RFC 1889, RTP: A Transport Protocol for Real-Time Applications, January 1996
8. IETF RFC 2326, RTSP: Real Time Streaming Protocol (RTSP), January 1996
9. D.F. McAllister (ed.), Stereo Computer Graphics and Other True 3-D Technologies, Princeton, NJ: Princeton University Press, 1993

An MPEG-4 Authoring System with Temporal Constraints for Interactive Scene*

Heesun Kim

Division of Computer & Multimedia Engineering, Uiduk University,
780-713, 525 Yugeom, Gangdong, Gyongju, Korea
kimhs@mail.uiduk.ac.kr

Abstract. An MPEG-4 scene is the specification for generating interactive multimedia contents. Each object constituting an MPEG-4 scene runs according to its own run time. Likewise, it should support the update of predefined temporal relations and attributes by the user event taking place during run time. Nonetheless, BIFS, which is the scene description of MPEG-4, does not support the temporal relations among the objects; neither is it capable of controlling the variance of temporal properties by user events. This paper defined the temporal relations and its related events that are helpful to the effective authoring of MPEG-4 and introduced temporal constraints to generate error-free scenes for user events taking place during run time. Furthermore, an authoring system of MPEG-4 contents was developed together with this application.

1 Introduction

MPEG-4 scene consists of various multimedia objects, with each object played at its own playing time. For efficient authoring of the audio/visual object that composes the MPEG-4 scene, it is important to author the playing time for each object and temporal relation [1–3]. However, the time model of MPEG-4 is based on time-stamped events and does not offer a method to express temporal dependence among audio/visual objects [4]. Likewise, the MPEG-4 scene is defined to handle the user event that occurs during the play. Therefore, it should handle the variation of playing time and relationship of each object by the user event that occurs during the play, as well as the temporal relation among the objects. Furthermore, constraint of user event production is needed for the object in which the temporal relation is established to maintain temporal consistency, even if the event happened already.

Regarding researches related to this research, BIFS (BInary Format for Scene) [5] is a scene description language of MPEG-4 and its authoring tool. BIFS of MPEG-4 is a scene format made up of audio/visual nodes such as audio, video, and 2D and 3D geometric body, property-representing node, and event-related node, etc. The node and field of BIFS were designed based on VRML [6]. The

* This work was supported by grant No.(R04-2002-000-20026-0) from the Basic Research Program of the Korea Science & Engineering Foundation.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 457–465, 2004.

playing time for each object in BIFS is based on time-stamped event, and each object's starting and ending time of play appear separately. That is, temporal relation with other objects and event-related time variation are not defined. For the MPEG-4 content authoring system, there are [Boughoufalah et al. 2000] [4], [Viljoen et al. 2002] [7]. These authoring tools offer visual authoring environment and write the time information among the objects using the time setting bar. They provide sequential and parallel play relationship as the object temporal relation but the handling play time according to the user event is not satisfactory.

This paper defines the playing time and relationship of the object to the user event for the MPEG-4 scene with dynamic characteristic, and creates an MPEG-4 content authoring system that supports an error-free scene play for this relationship setting. The proposed temporal constraint supports the playing order, time, and time variation by the user event during the play. Likewise, the constraint about temporal relation and event setting is defined to give consistency about the written temporal relation. Proposed authoring system composes the scene and it is easy to use for common users without knowledge about MPEG-4 as it offers a visual authoring environment, and can write time and event. Chapter 2 discusses the temporal constraints of the making temporal relations among audio/visual objects and user interaction. Chapter 3 shows the MPEG-4 content authoring system that supports the temporal relations. Chapter 4 illustrates the development and evaluation authoring system. Lastly, Chapter 5 gives the conclusion.

2 The Definition of Time-Related Event and Temporal Constraints

2.1 Temporal Relations and Constraints

The temporal relation for each object comprising the MPEG-4 Scene is defined by analyzing the MPEG-4 systems. The temporal relation Tr is defined as follows.

Tr={Equal,Overlap,During,CoStart,CoEnd,Sequence,After,Exclusive}
This paper considered various cases for ensuring flawless contents when setting up temporal relations and events. Table 1 shows the constraints for the temporal relation, which can be established when making the MPEG-4 scene. m_i and m_j in Table 1 represents the arbitrary audio-visual relationship. Δt represents the time segment greater than 0. If the audio-visual object comprising a scene as m is represented, the time attribute of m is defined as follows.

$$\mathrm{m} = (\mathrm{m.s},\ \mathrm{m.e},\ \mathrm{m.d},\ \mathrm{m.md})$$

m.s represents the start time of the play of the object and m.e represents the end time of object. m.d represents the play continuation time and m.md represents the maximum continuation time. Sequence relationship requires that end time of m_i coincides with the start time of m_j, while the Costart relationship requires that the start time of m_i should coincide with the start of m_j. CoEnd relationship

Table 1. Temporal relations and constraints

Temporal relations	Constraints
Sequence(m_i, m_j)	m_i.e = m_j.s
CoStart(m_i, m_j)	m_i.s = m_j.s
CoEnd(m_i, m_j)	m_i.e = m_j.e
Equal(m_i, m_j)	(m_i.s = m_j.s) $\wedge$ (m_i.e = m_j.e)
Overlap(m_i, m_j)	(m_j.s - m_i.e $\leq \Delta$t) $\wedge$ (m_i.s - m_j.s $\leq \Delta$t) $\wedge$ (m_i.e - m_j.e $\leq \Delta$t)
During(m_i, m_j)	(m_i. s - m_j.s $\leq \Delta$t) $\wedge$ (m_j.e - m_i.e $\leq \Delta$t)
After(m_i, m_j)	m_i.e - m_j.s $\leq \Delta$t
Exclusive(m_i, m_j)	(m_i.s = m_j.s) $\wedge$ (m_i.e = m_j.e)

requires that the end of m_i should coincide with m_j, while Equal relationship requires that the start and end of m_i should coincide with those of m_j. Overlap relationship requires that the start of m_i is earlier than the start of m_j and the end of m_i is later than the start of m_j. During relationship requires that the start of m_i play is earlier than the start of m_j play and end of m_i play is later than the end of m_j play. After Relationship requires that the end of m_i be earlier than the start of m_j. The Exclusive relationship requires the same constraints as the Equal relationship.

2.2 Time-Related Events and Constraints

The time-related event E established for the MPEG-4 scene is as follows.

E={eStart, eStop, eScale, eOrder, eChangeTr, eHyperLink}

eStart event determines the start of the object play dynamically according to the user events. eStop event determines the end time of an object dynamically according to user events. eScale event extends or shortens the play continuation of an object. eOrder event alternates the order of object play. eChangeTr event changes the static temporal relation established for an object to another temporal relation. ehyperLink is an event to change a scene to another that is connected to it. The following are the handling methods for each event.

(1) eStart event

Investigates if the object to which an event is established has temporal relation with an-other object. If it has temporal relation with another, it finds all objects that have direct or indirect relationship with it. All related objects are affected by the eStart event and the temporal relation established earlier is kept after recalculating the playing time from the event start time.

(2) eStop event

If eStop event happens for an object, the corresponding event ends. If there is an object, which has a temporal relation with the terminated object, it recalculates the playtime since the event happens and keeps the temporal relation established earlier.

(3) eScale evernt

If an event happens, the object playtime is changed. The play continuation time of the changed object, md, should be greater than 0 and smaller than the object's maximum play continuation time, m.md.

$$0 < \mathrm{m.d} < \mathrm{m.md}$$

(4) eOrder eventt

The constraint for establishing events is the temporal relation, tr_i has After relationship with Sequence.

$$tr_i \in \{ \text{ Sequence, After } \}$$

(5) eHyperLink event

If an event happens, the object played earlier is terminated and the connected object is played.

(6) eChangeTr event

Since the object playtime is not changed for the object whose temporal relation is already established, it cannot be changed regardless of the constraints, while in some cases it can be changed regardless of constraints. Table 2 lists the constraints for the changeable temporal relation when the constraint is satisfied.

In Table 2, m_i and m_j are arbitrary audio visual object different to each other and unit of Δt is second. If the relationship of m_i and m_j is changed from Sequence, After or Overlap to Equal or Exclusive, the constraint that the play continuation time of m_i and m_j should be equal has to be satisfied. If the relationship of m_i and m_j is changed from Sequence, CoStart, CoEnd, After or Overlap to During, the constraint that the play continuation time of m_i and m_j should be equal has to be satisfied.

Table 2. Constraints according to the change of temporal relation

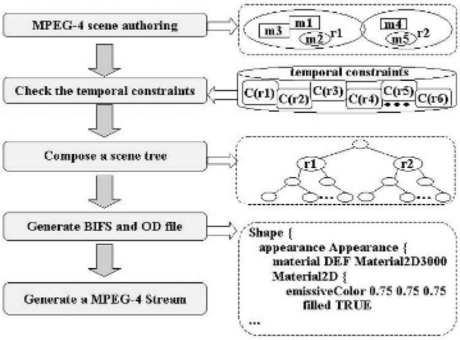
Relationship change	Constraints
Sequence(m_i , m_j) → Equal(m_i , m_j),Exclusive(m_i , m_j)	m_i.d = m_j.d
Sequence(m_i , m_j) → During(m_i , m_j)	m_j.d - m_i.d $\leq \Delta$t
CoStart(m_i , m_j) → During(m_i , m_j)	m_j.d - m_i.d $\leq \Delta$t
CoEnd(m_i , m_j) → During(m_i , m_j)	m_j.d - m_i.d $\leq \Delta$t
After(m_i , m_j) → Equal(m_i , m_j), Exclusive(m_i , m_j)	m_i.d = m_j.d
After(m_i , m_j) → During(m_i , m_j)	m_j.d - m_i.d $\leq \Delta$t
Overlap(m_i , m_j) → Equal(m_i , m_j), Exclusive(m_i , m_j)	m_i.d = m_j.d
Overlap(m_i , m_j) → During(m_i , m_j)	m_j.d - m_i.d $\leq \Delta$t

3 MPEG-4 Scene Authoring System

This chapter states the MPEG-4 authoring system supports the time-related event authoring and suggests the authoring environment for establishing a temporal relation, a method for generating scene composition tree and a method for generating written scenes with MPEG-4 stream.

3.1 Generating MPEG-4 Scene

This MPEG-4 authoring system enables users to make contents easily, fast and effectively by providing visual and intuitive user interface. The process for generating MPEG-4 stream by writing the MPEG-4 Scene is shown in Figure 1.



```
Shape {
  appearance Appearance {
    material DEF Material2D3000
    Material2D {
      emissiveColor 0.75 0.75 0.75
        filled TRUE
...
```

Fig. 1. MPEG-4 Scene generation Process

Users can author the audio-visual object, temporal relation, route, and command information visually through the user interface. Inspecting the validity of temporal relation established between objects reflects the time information on the scene tree if the relationship is valid. Otherwise, the author is notified that the relationship is not valid. The scene tree is composed of the audio-visual object written in the user interface with temporal relations, event information, and attributes. The scene tree is the data structure used in this system. The information for generating BIFS and OD(Object Descriptor) is extracted by searching the scene tree and the BIFS, and the OD file is then generated. The BIFS, OD, and Stream are composed and encoded to MPEG-4 stream.

3.2 Generation of the Scene Composition Tree Using Time Constraint Inspection

The suggested event supporting the time constraint model is represented as a scene tree in a system. N is defined as the set of nodes making up a tree. The kinds of node comprising a tree are as follows.

```
N = { Nm, Ntr, Ne, Np }
Nm : Set of nodes representing audio-visual objects
Ntr : Set of nodes representing temporal relations
Ne : Set of nodes representing time-related events
Np : Set of nodes representing attribute information
     of audio-visual objects
```

For the scene composition tree suggested in this paper, the event node, temporal relation node, and audio-visual node can be a child object of a group object with the group object as its root. If the temporal relation between audio-visual objects is not established, the audio-visual object is connected to the group object. The latter is connected to the object with temporal relation under its node. If an event is established to an object with temporal relation, the event node is placed over its node. The audio-visual object again has an attribute object. Figure 2 shows the scene composition tree generation process for establishing the time-related event.

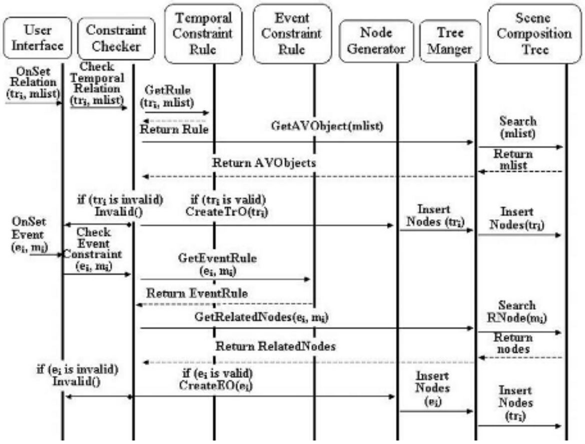


Fig. 2. Scene composition tree generation process for establishing temporal relation and events.

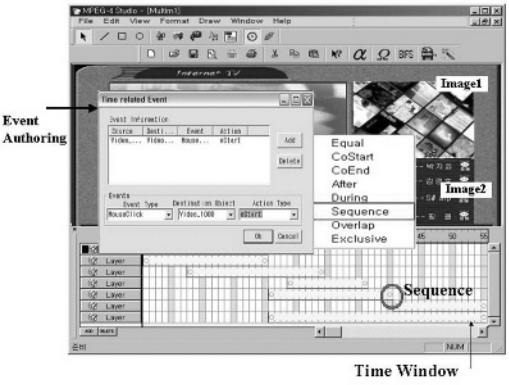
If the temporal relation tr_i is established to the audio-visual object list mlist in the user interface, the Constraint Checker inspects the time constraints with time attributes included in the mlist object referencing the rules for time constraints. If the established temporal relation satisfies the constraints, a time-related node is generated and inserted in the scene composition tree. If event ei is established to the audio-visual object m_i, it obtains the event-setting rule and examines if the audio-visual object m_i has a temporal relation. If m_i has a temporal relation, all the objects related to m_i will have both direct and indirect temporal relations. ei is established in all the objects with temporal relation with m_i; if they are found to be valid, they are inserted in the scene composition tree. The event node is inserted as a parent node of the time-related node.

3.3 MPEG-4 Scene Generation and Stream Generation

In the MPEG-4 scene of the BIFS and OD text file, searching the scene composition tree and referencing the BIFS generation rule of MPEG-4, OD rule, route rule, and command rule generates the OD of an object information. The scene information is obtained from the scene composition tree, which generates the BIFS text file. The scene tree is searched from the root node, and each node generation rule is referenced to write the corresponding node as a BIFS text. The generated BIFS and OD files are encoded as binary type, and the encoded BIFS, OD, and media files are incorporated into the MPEG-4 Stream. At this time, they are encoded into the MPEG-4 stream referencing the NCT(Node Coding Table), NHI(Node Hierarchy Information), and NDT(Node Data-type Table).

4 Development and Evaluation

This authoring system is developed using MS-Windows XP and Visual C++6.0. The MPEG-4 content authoring system is based on a visual authoring environment. The playing time of each object can be established using a time window, and the temporal relation can be set in the pop-up menu after selecting the objects for establishing the temporal relation. The time window helps the author grasp the whole play scenario. It also updates the related objects automatically when changing the object's playing time with the temporal relation and constraints. Figure 3 is an example of this authoring system establishing temporal relation and event in audio-visual objects. The playing time of Image1 and Image2 is set in a time window, while the Sequence relationship is set using the pop-up menu. For establishing a time-related object, the kind of event, target object, and action is selected in the event definition dialog box after selecting an object. The target object refers to the objects itself and objects comprising the contents, whereas the action represents the time-related event defined in this paper. Figure 4 shows the play of contents where the temporal relation and event



Event Information

Source	Desti...	Event	Action
Video,...	Video...	Mouse...	eStart

Events

Event Type: MouseClick
Destination Object: Video_1000
Action Type: eStart

Fig. 3. Temporal relation and event authoring example

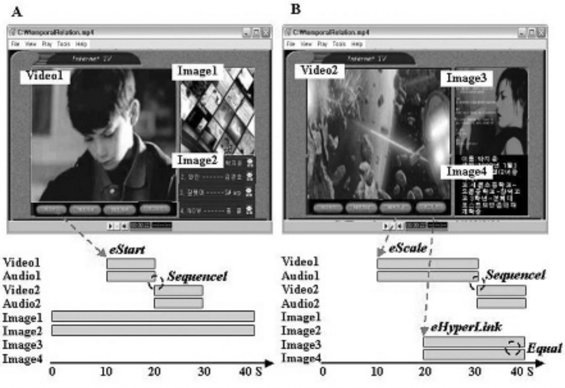


Fig. 4. An example of playing MPEG-4 contents

are authored. The contents comprising object has various temporal relations and time-related object The validity of the authoring system is verified by examining if the content is played as intended when time passes or user interaction occurs. A part in Figure 4 shows the scene when the time is 12 seconds after the start. Video1 and Audio1 have the Equal relationship, and the eStart event is set to Video 1. If a user clicks the image button 10 seconds after the start, Video1 is played by the eStart event, and the Audio event with Equal relationship with Video1 is played simultaneously. Accordingly, the temporal relation and event set to Video1 is played with validity. The problem of breaking the Equal relationship when Video1 has the Equal relationship and eStart event simultaneously occurs for the authoring tools with no time-related events handling scheme. B part in figure 4 shows the scene 32 seconds after the start. The eHyperlink event occurs 20 seconds after the start, making Image1 disappear and playing Image3. In this case, Image2 with the Equal relationship with Image1 disappears and Image4 with the Equal relationship with Image3 is played.

This authoring system reduces the number of time authoring operation of each object to support authoring the effective temporal relation. If a playing time of an object is changed, the playing time of related objects are changed automatically for the authoring systems supporting temporal relation. If there is no temporal relation, however, the playing time of each related object should be changed manually. In the absence of the temporal relation, the n^2 authoring operation is required for the n time changes. In case of supporting temporal relation, however, the n authoring operation is required for the n time changes, which enables fast and easy authoring. Furthermore, this authoring system provides valid temporal relation and has the advantage of automatic update of related object time according to the constraints in case the playing time changes the object by defining both the temporal relation for each object and the constraints that the relationship should satisfy.

5 Conclusion

This paper suggests the temporal relations and constraints for interactive MPEG-4 scene authoring. The temporal relations that set the objects making up the MPEG-4 scene and events are defined, and the constraints and event-handling method are described. These temporal relations and MPEG-4 contents with user event will have time attribute changing dynamically by the user event occurring during play. And it will be played as an error-free content according to the author's intention. The MPEG-4 contents authoring system is developed based on these temporal constraints model. This system provides intuitive and visual authoring environment to enable users with no knowledge of the MPEG-4 scene constitution to author contents easily. This MPEG-4 content authoring system provides temporal relation authoring environment, constraint inspection, scene composition tree generation for the established temporal relation, and MPEG-4 stream generation according to the MPEG-4 generation rule. This time constraint model and content authoring system reflect user interaction fully to overcome the limits of expressing the temporal relation of MPEG-4 contents.

References

1. J. F. Allen, "Maintaining Knowledge about Temporal Intervals," Communications of the ACM, vol. 26, no. 11, pp. 832–843, 1983
2. K.Cha and S.Kim, "Authoring Temporal Scenarios in Interactive MPEG-4 Contents," Lecture Note of Computer Science(LNCS), Springer-Verlag : Proceedings of the 3th IEEE Pacific-Rim Conference on Multimedia, pp. 1235–1242, 2002
3. B. Prabhakaran and S. V. Raghavan, "Synchronization Models for Multimedia Presentation with User Participation," Proceedings of the First ACM International Conference on Multimedia, pp. 157–163, 1993
4. S. Boughoufalah, J. Dufourd and F. Bouihaguet, "MPEG-Pro, an Authoring System for MPEG-4 with Temporal Constraints and Template Guided Editing," Proceedings of the 2000 IEEE International Conference on Multimedia and Expo, 2000
5. Information Technology-Coding of Audio-Visual Objects-Part 1 : Systems, ISO/IEC 14496-1, ISO/IEC JTC 1/SC 29/WG 11, 1998
6. VRML 97, ISO/IEC DIS 14772-1, 1997
7. D.W. Viljoen, A.P.Calitz, N.L.O.Cowley, "A 2-D MPEG-4 Multimedia Authoring Tool," Proceedings of the 2nd international conference on Computer graphics, virtual Reality, visualisation and interaction, 2003

A Method of Digital Camera Work Focused on Players and a Ball

– Toward Automatic Contents Production System of Commentary Soccer Video by Digital Shooting –

Masahito Kumano[1], Yasuo Ariki[2], and Kiyoshi Tsukada[3]

[1] Faculty of Science and Technology, Ryukoku University,
kumano@rins.ryukoku.ac.jp
[2] Faculty of Engineering, Kobe University,
ariki@kobe-u.ac.jp
[3] Mainichi Broadcasting System, Inc.
tsukada@mbs.co.jp

Abstract. We aim at the realization of automatic adaptive contents production system of commentary soccer video based on digital shooting technique which is composed of digital camera work and digital switching technique for small audience such as a fellow some enthusiast and a individual. The digital camera work is defined as virtual panning and virtual zooming. Also, the digital switching is defined as the change of some virtual camera by controlling rapid change of frame location or size on a HD(high definition) material video. In this paper, we describe a method of digital camera work technique focused on players and a ball as a sub-system of automatic contents production system for commentary soccer video. The produced video contents with digital camera work is subjectively evaluated by AHP (Analytic Hierarchy Process).

1 Introduction

In the coming digital age, both a lack of video contents and a prodigious amounts of work to new interactive services make a serious problem because of increasing number of digital broadcast channels. Therefore a large quantity of broadcast contents are strongly required. To such a problem, private sport video contents, which do not necessarily need professionality of an editor, a switcher and a camera man, is a key issue to supply for the small audience. The private sport video contents, however, have not been produced so far because of the production cost. Therefore, the efficient and automatic content production system is required at low cost without professionality.

In sport video contents, the soccer game attracts a global viewership. However, it is said that soccer fans cannot grow up easily in Japan, because process of soccer game is unclear for amateur about soccer. To solve these problems, we aim at the realization of an automatic contents production system of commentary soccer video based on a digital shooting technique which is composed

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 466–473, 2004.

of digital camera work and digital switching technique. In sports live contents, zooming is seldom used except for fine adjustment of a frame size. Therefore, in this paper, we focus on the realization of the digital panning with the fixed frame size. Although various realization of panning mode can be assumed, we follow the same camera work used in TV as a foremost task in this paper.

There are many approaches to the contents production system such as generating highlight [1,2], summary [3,4], reconstruction [5], mosaic [6] and these elemental technology [7] to sport TV program contents. However, these produced contents are greatly subjected to restriction of contents production staff such as a cameraman, editor, switcher, etc. Also, these contents inherit mistake of camerawork and switching. Therefore, these contents have little degrees of freedom. The digital shooting has the degrees of freedom higher than secondary usage of sport TV program contents, because it is able to generate variant camera work and switching from material video contents taking a whole soccer court by HD camera. As a related work in automatic retakable shooting system that we call "one source multi-production system", Virtual Soccer Stadium [8] shows a free view in 3D space of soccer court to viewers. However it is another matter where is shown to viewers. In other words, the soccer video contents by TV program maker using camera work or switching teach viewers where to watch as a primary commentater.

In this paper, we describe two control methods of clipping frame in digital camera work focusing on players and a ball as a elemental technology of contents production system for commentary soccer video. In Section 2, the digital shooting is described. In Section 3, experiments of digital camera work focused on the players are presented. In Section 4, experiments of digital camera work focused on the ball are presented. In Section 5, we uses AHP, in order to carry out subjectivity evaluation of the produced video contents with digital camera work.

2 Digital Shooting

In a production of sports live contents by work of cameraman and switcher, a multiple camera system is used in filming. The digital shooting can be assumed as an emulation of a virtual multiple camera system by clipping the frame from HD material video contents taking a whole coverage area by HD camera and by mapping roughly to frame with the resolution for example SD(Standard Definition). The digital shooting technique is composed of digital camera work and digital switching technique. The digital camera work is defined as virtual panning and virtual zooming. the virtual panning is a video production technique of clipping a size-fixed frame by controlling frame location on a HD material video. The virtual zooming is a video production technique of clipping a frame by controlling frame size. Also, the digital switching is defined as the change of some virtual camera by controlling rapid change of frame location or size on a HD material video.

Although the camera work or switching by human in live sport cannot perform retaking due to environmental cause, the digital shooting is able to repeat-

Table 1. Standard and Resolution of High Definition

	Standard	Screen resolution ratio	Screen resolution
1	SD D1:525i,D2:525p	9:16,3:4	720x480
2	HD D4:750p	9:16	1280x720
3	HD D3:1125i,D5:1125p	9:16	1920x1080
4	Next	9:16	3840x2160

edly produce various camera work and switching from material video contents taking a whole soccer court by HD camera. Therefore the digital shooting production system can perform as potential ability the various virtual taking and meet request of small audience.

In the Digital Shooting, the material video is basically be taken by fixed HD camera. Therefore the material video does not include the camera work by human. So the background subtraction method apply to the material video to extract players and ball. The background subtraction is a simple but effective method to detect moving objects in video images. These background subtraction images do not depend on the image color to extract moving objects, hence this method can apply not only to the grass court but also to the earthen court and correspond with slow transition of sunshine or illumination by updating these background subtraction images. However, the material video must include the whole of soccer court seamlessly for the real system. About this problem, the seamless material video can be realized technically because panorama video contents production system using three HD camera(D5 in the Table 1) [9] is known. In this paper, we used the material video contents including the half-court taken by one HD camera(D3 in the Table 1).

In the background subtraction process, each binarized images is preprocessed by a morphological operator (erosion and dilation) to extract a region of player, and is applied noise reduction processing. Then, the successive regions temporally are defined as the moving object.

3 Frame Position Control Focused on Players

In this section, we present a method to decide panning mode by using moving information of soccer players. A shooting method as an emulation of TV is considered to locate the center of the soccer game process. In the background subtraction images, moving object such as player can be extracted easily. In process of the soccer game, however, moving objects exist in the whole of the soccer court. Then, we focused on the area of those players crowding and moving quickly near the ball. First, $k-th$ binarized background subtraction image is defined as f^k. In f^k, N moving objects are labeled to identify such as O_k^1,O_k^2,O_k^3,$\cdots$,O_k^n,$\cdots$,O_k^{N-1},O_k^N. The location ($x-y$ coordinate in the background subtraction image) of a player is $\mathbf{P}^{(O_k^n)}(P_x^{(O_k^n)}, P_y^{(O_k^n)})$ computed as a centroid of $n-th$ moving object using the Eq. (1), and $\mathbf{P}^{(O_k^n)}$

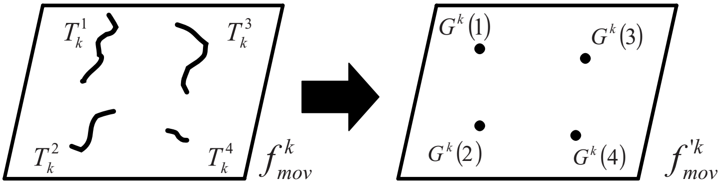


Fig. 1. Process of extracting centroid of moving objects

generates a trajectory T_k^n by plotting iteratively within the proceeding frames $f^{k+l}(l = 0, 1, 2, \cdots, L : L = 59)$, resulting in the moving information image f^k_{mov} (the left side in Figure 1) computed by Eq. (2).

$$P_x^{(O_k^n)} = \frac{\sum_{x \in O_k^n} x}{\sum_{x \in O_k^n} 1} \qquad P_y^{(O_k^n)} = \frac{\sum_{y \in O_k^n} y}{\sum_{y \in O_k^n} 1} \tag{1}$$

$$f^k_{mov} = \begin{cases} 255 \;(\; \mathbf{P}^{(O_k^n)} \neq \mathbf{P}^{(O_{k+l}^n)}, (l = 1, 2, 3, \cdots, L) \;) \\ 0 \quad\;\; (\; \mathbf{P}^{(O_k^n)} = \mathbf{P}^{(O_{k+l}^n)}, (l = 1, 2, 3, \cdots, L) \;) \end{cases} \tag{2}$$

Second, a trajectory centroid $G^k(n) = (G^k_x(n), G^k_y(n))$ is computed to each T_k^n (the right side in Figure 1) and the weight $W_k(n) = s_n$ is computed as the number of pixels s_n included in each T_k^n. Here, the player's speed is regarded high if $W_k(n)$ has high value because it can be assumed that a longer trajectory means a larger movement of a player. Third, the frame location control point of panning is defined as a most crowded location. Then, after trajectory centroids $G^k(n)$ of all moving objects O_k^n are computed, we approximate a most crowded location $G^k_c(n_k)$ as shown in Eq.+(5) as a location of n_k computed by Eq.+(4).

$$W'_k(n) = \frac{W_k(n)}{W_{k,total}}, \quad W_{k,total} = \sum_{n \in N} W_k(n) \tag{3}$$

$$n_k = \arg\max_n \sum_{t \in N, t \neq n} \frac{W'_k(t)}{\{G^k_x(n) - G^k_x(t)\}^2 + \{G^k_y(n) - G^k_y(t)\}^2} \tag{4}$$

$$G^k_c(n_k) = G^k(n_k) = (G^k_x(n_k),\; G^k_y(n_k)) \tag{5}$$

Finally, the panning trajectory is computed as a linear regression line of $\mathbf{G}_c^{k+m}(n_k)$ $(m = 0, 1, 2, \cdots, M : M = 59)$ for each half of $M + 1$ frames in order to get the smooth motion of panning. The parameter α and β of the linear regression line $y = \alpha x + \beta$ is computed using Eq.(6), Eq.(7) and the amount of panning movement (x_m, y_m) per one frame is computed by Eq. (8).

$$\bar{G}_x(n_k) = \frac{1}{M+1} \sum_{l=0}^{M} G_x^{k+l}(n_k) \qquad \bar{G}_y(n_k) = \frac{1}{M+1} \sum_{l=0}^{M} G_y^{k+l}(n_k) \tag{6}$$

$$\alpha = \frac{\sum_{l=0}^{n}(G_x^{k+l}(n_k) - \bar{G}_x(n_k))(G_y^{k+l}(n_k) - \bar{G}_y(n_k))}{\sum_{l=0}^{n}(G_x^{k+l}(n_k) - \bar{G}_x(n_k))^2}, \ \beta = \bar{G}_y(n_k) - \alpha\bar{G}_x(n_k) \tag{7}$$

$$x_m = \frac{2(G_y^{k+(M+1)/2}(n_k) - G_y^k(n_k))}{\alpha(M+1)}, y_m = \frac{2\alpha(G_x^{k+(M+1)/2}(n_k) - G_x^k(n_k))}{(M+1)} \tag{8}$$

4 Frame Position Control Focused on a Ball

In this section, we present method to decide panning mode using moving information of a soccer ball. The ball location is important because the soccer game progresses following a ball location. However, the ball trajectory is not adequate to use directly for generating the panning trajectory because a smooth panning cannot be obtained due to speedy and wiggled movement of the ball. In other words, the panning mode have to satisfy two contrary essence that the frame follows the ball quickly if the ball moves widely and speedily, and if the ball moves wiggly and speedily, the frame follows slowly or not follows. To solve this problem, we use inside frame such a black frame shown in Figure 2, provided we index the trajectory of the ball by manual to focus on the control method of the clipping frame in this paper.

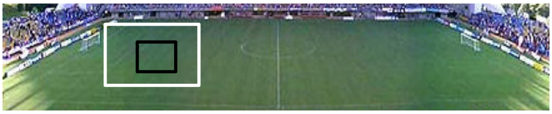


Fig. 2. Frame (white) and Inside frame (black)

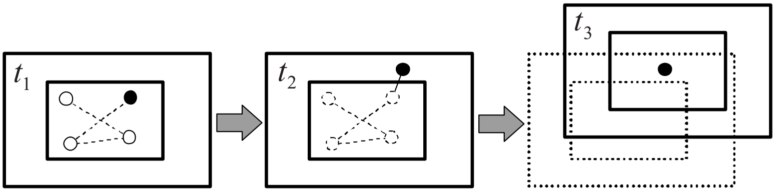


Fig. 3. Methods of frame control

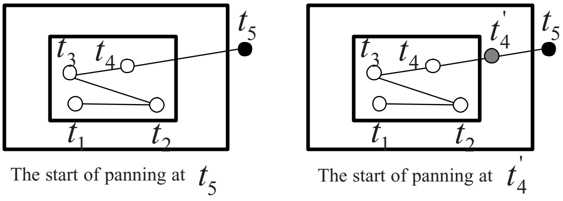


Fig. 4. Methods of frame extended control by forecast vector

For wiggled movement, the frame stills if the ball exists inside the black frame, and if the ball is out of the black frame, the centroid of white frame moves to the ball location as shown in Figure 3. Also for wide and speedy movement, if the ball is just out of the black frame wildly as shown at t_5 in the left of Figure 4, the frame will move hurriedly. In order to predict a big motion of the ball and to make the frame controller react early, we employ a forecasted vector $V = (V_x, V_y)$ of a moving ball computed between current and previous frame to forecast the wide movement of the ball in the black frame. In the proposed method, at the current time t_4, if a forecast vector shown at t'_4 in the right of Figure 4 is out of the black frame, the white frame moves so that the ball exists in the black frame. By this method, the frame can move smoothly and early to the wide movement of the ball.

5 Evaluation Experiment

5.1 Subjective Evaluation by AHP

In this section, we present a method to evaluate the produced video contents. Our final goal of this study is subjectively to realize the commentary soccer contents production system which enables preferences of small audience such as fellow-enthusiast and a person appropriately. Therefore, AHP (Analytic Hierarchy Process) [10] is used for the evaluation of the produced video contents because of its qualifications to represent human's subjectivity. The AHP is a multicriteria decision support method designed to select the best from a number of alternatives evaluated with respect to several criteria. It carries out pairwise comparison judgements which are used to develop overall priorities for ranking the alternatives.

5.2 Evaluation Experiment by AHP

The criteria of evaluation is based on understandability related to the camera work. Therefore, the panning trajectory and speed are adopted as the criteria of naturality. The video quality is adopted to judge whether the produced video contents pale against TV or HD contents or not. Also, understandability of game process is adopted additionally. They are shown at middle layer of AHP treemap in the upper of Figure 5. Here, the produced contents described in section 3 is called ‘contents of experiment 1’, and one described in section 4 is called ‘contents of experiment 2’. Also, TV contents and HD contents of soccer game are compared as alternatives with the contents of experiment 1 and experiment 2. The reason why the HD contents are compared is to investigate which is fundamentally more comprehensible between TV contents with camera work and HD contents taking wide-angle of soccer court. These alternatives are shown at bottom layer of AHP treemap in the upper of Figure 5

All of these alternatives at bottom layer of AHP treemap in the upper of Figure 5 are produced from the same soccer game of 38th National high school

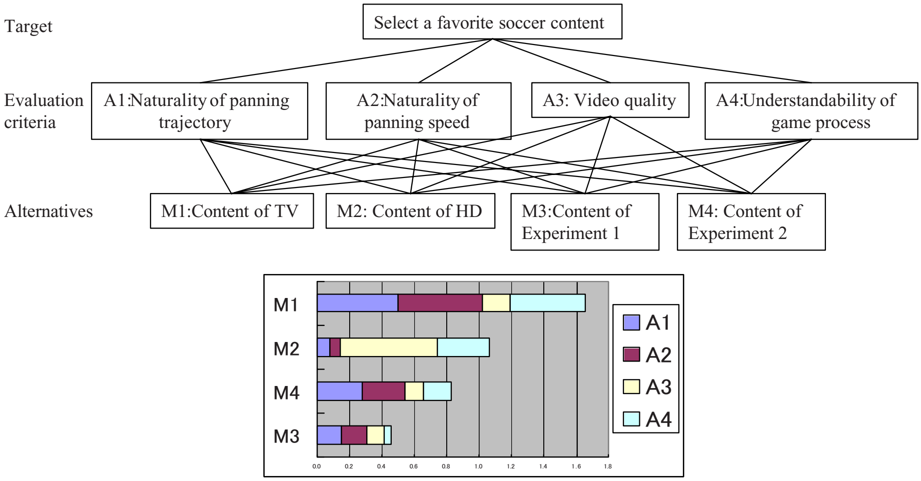


Fig. 5. AHP treemap (upper) and Experimental result (lower)

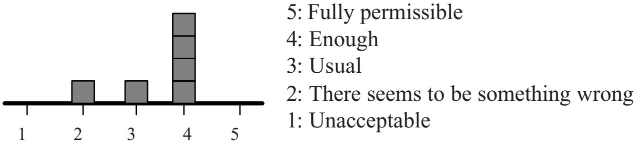


Fig. 6. Result of tolerance questionnaire

soccer championship Kyoto area final in Japan. The HD contents is taken for a wide-angle half of the court by HD camera (D4 level as shown in Figure 1) because of limitation of the camera in this paper. The experiment was conducted by showing the contents composed of 6 shot including pannings to six examinee.

As a result of the experiment, the middle layer of AHP showed the preference weights A1,A2,A3,A4 = 0.15, 0.12, 0.25, 0.48. This result indicates that the understandability is the most important criterion and then the resolution(video quality) follows. Also, the weights of AHP is shown in descending order at the bottom of Figure 5. This result shows that the presentation of game process is the most important and therefore the TV content was selected. In fact, it indicates that the camera work by shooting techniques of camera man is important compared to no camera work. On the other hand, the contents produced by the proposed method is inferior to the TV contents because it is assumed that the shooting high techniques by professional camera man is not realized.

5.3 Result of Tolerance Questionnaire

In addition, we carried out tolerance questionnaire as the private soccer video contents. The result shown in Figure 6 indicates that the contents produced by the methods described in Section 3 and Section 4 are enough as the private soccer video which does not necessarily need professionality. As a result, although the contents produced by the proposed method is inferior to TV contents or HD contents, they are accepted for the private soccer video contents which do not necessarily need professionality.

6 Conclusion

In this paper, a method was proposed to realize an automatic digital panning forcussing on moving players or a ball. Future works will be not only the improvement of the digital camera work to emulate professional camera man, but also realizing the digital shooting and research of a new way to show the soccer game process intelligibly.

References

1. Dennis Yow,Boon-Lock Yeo,Minerva Yeung,Bede Liu: “Analysis and Presentation of Soccer Highlights from Digital Video”, ACCV’95, pp. 499–503.
2. Jurgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, Walter Nunziati: “Automatic Interpretation of Soccer Video for Highlights Extraction and Annotation”, SAC 2003: pp. 769–773
3. A. Ekin, A. M. Tekalp, and R. Mehrotra: “Automatic soccer video analysis and summarization”, IEEE Trans. on Image Processing, vol. 12, no. 7, pp. 796–807, July 2003
4. Ngoc Thanh Nguyen, Tuoung Cong Thang, Tae Meon Bae, Yong Man Ro: “Soccer Video Summarization System Based on Hidden Markov Model with Multiple MPEG-7 Descriptors”, CISST 2003: pp. 673–678.
5. Thomas Bebie, Hanspeter Bieri: “SoccerMan – Reconstructing Soccer Games from Video Sequences”. ICIP (1) 1998: pp. 898–902
6. Hyunwoo Kim, Ki-Sang Hong: “Soccer Video Mosaicing Using Self-Calibration and Line Tracking”. ICPR 2000: pp. 1592–1595
7. Okihisa Utsumi, Koichi Miura, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka: “An object detection method for describing soccer games from video”, Proc. 2002 IEEE Intl. Conf. on Multimedia and Expo (ICME2002), vol. 1, pp. 45–48 (Aug. 2002)
8. Takayoshi Koyama, Itaru Kitahara, Yuichi Ohta: “Live Mixed-Reality 3D Video in Soccer Stadium”, ISMAR 2003: pp. 178–187
9. http://www.megavision.co.jp/
10. Saaty, T.: “A scaling method for priorities in hierarchical structures”, Journal of Mathematical Psychology, Vol. 15, pp. 234–281 (1997)

Real-Time Rendering of Watercolor Effects for Virtual Environments

Su Ian Eugene Lei and Chun-Fa Chang

Department of Computer Science, National Tsing Hua University
{zenith,chang}@ibr.cs.nthu.edu.tw

Abstract. In this paper, we present an empirical approach for rendering realistic watercolor effects in real-time. While watercolor being a versatile media, several characteristics of its effects have been categorized in the past. We describe an approach to recreate these effects using the Kubelka-Munk compositing model and the Sobel filter. Using modern per-pixel shading hardware, we present a method to render these effects in an interactive frame-rate.

Keywords: Non-photorealistic rendering, graphics hardware, image processing, Kubelka-Munk, watercolor.

1 Introduction

While non-photorealistic rendering (NPR) is a relatively new field, many different approaches for variant styles have been proposed through the past decade. With increasing computational power and improving graphics hardware architecture, a number of approaches on implementing those effects in real-time has been proposed. So far, the real-time automation scene of the NPR field mainly focuses on simulating stroke-based effects, such as pencil hatching [14], pen-and-ink [15], engraving [6] etc. Medium that requires extensive cell-to-cell interaction of pigments, such as watercolor, is difficult to render in real-time realistically.

In this paper, we followed the work of Curtis et al. [2], which characterizes several defining effects of watercolor. We propose a system that can render these effects in real-time, with the assistance of per-pixel shading hardware. We also demonstrate the use of Sobel filter for simulating edge darkening and granulation characteristics of watercolor.

1.1 Related Work

The pioneering work of Small [1] used a Connection Machine to simulate watercolor. Curtis et al. [2] used a similar approach, which used an improved physical simulation model to achieve more realistic effects. Their method, while looks realistic and physically sound, requires too much computation to achieve an interactive frame-rate. Inspired by the work of Lake et al. [5], we use a "color-band" approach instead of simulating fluid and pigment interaction for each

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 474–481, 2004.

frame. Such color-band is defined by using a simplified Lit-sphere [4] interface. We use a pigment-blending model based on the Kubelka-Munk [3] (KM) model to simulate the composition of different pigments.

One of the most obvious effects of watercolor is edge darkening, i.e. a dark deposit at the edge of a wet-on-dry stroke. Intuitively, recreating this effect as a post-process is similar to the silhouette-finding problem. While silhouette sketching is a well-understood field [9,13], conventional methods focus on finding the silhouette of a geometric model. Instead we use the Sobel filter [18] to find the edge of every painted region efficiently. The work of Nienhaus et al. [8] describes a possibility to use 2D image processing process to enhance the results in real-time, and we implement our Sobel filter in a similar approach.

Recently some commercial applications and games used the Kuwahara filter [16] to create a watercolor style NPR. While this technique is fast, it fails to recreate the rich appearance of watercolor paintings, which depends on the KM model.

1.2 Organization

The rest of this paper is organized as follows. The next section reviews the characteristics of watercolor, as described by Curtis et al., and provides an overview of our approach. In section 3 we describe the system details of our implementation. We present the results and performance evaluation in Section 4.

2 Overview

Curtis et al. categorized several distinctive effects of watercolor: edge darkening, backruns, granulation, flow effects and glazing. Since our implementation is an automated system for creating artistic imagery, there are a few assumptions on when and where to apply such effects.

2.1 Watercolor Effects

Edge darkening is the key effect that most artists rely upon, and one of the most defining characteristics of watercolor. It is created when the pigment migrates from the interior of a wet region towards its edges as the paint begins to dry, leaving a dark deposit at the edge. We treat the entire painted area as the wet region, and add the darken edge by applying the Sobel filter.

Granulation of pigment creates a grainy texture concentrated on the peaks and valleys in the paper. The amount of granulation varies from paper to paper. We simulate the granulation effect by specifying a granulation constant for each kind of pigment, and use the paper texture and the Sobel filter to emphasize this effect.

Backrun is the effect when water is added to a wet painted area, carrying pigment along outwards, leaving a darkened, branched shape. Flow effect is observed when wet paint is applied on wet paper. The wet surface allows the pigment to spread freely, leaving a soft branching pattern. Color glazing is the process of adding thin layers of watercolor, causing pigments to be mixed optically.

We treat the backrun, flow effect and glazing as a universal effect. We simulate these effects by using a "color-band" approach. The color-band is a one-dimensional texture, which is created by mixing different amount of pigments in user-specified position. The color-band is then mapped onto the geometry using Lake et al.'s cartoon shading [5]. Additional flow effect is simulated in real-time using pixel-shader hardware.

2.2 Approach

Our system is divided into two phases: *Color-band specifying* and *watercolor shader*. The *color-band specifying* phase lets the user create a color-band for each object in the scene, using an isotropic lit-sphere interface [4]. The lit-sphere interface is set on top of a watercolor simulation engine, created using Curtis et al.'s water-flowing model [2]. In this stage, the flowing effect, backrunning and glazing are simulated as the user literally paints in the lit-sphere interface.

The *watercolor shader* is the main phase in our automated system. Using per-pixel shading hardware, we simulate edge darkening, granulation, paper texture and further flowing effects using shader programs. Details of the hardware-shader are discussed in section 3. A diagram is shown in Figure 1 to illustrate the various stages in our system.

Color-band Specifying. As mentioned in section 1, we use the Kubelka-Munk (KM) model [3] to perform the optical mixing of pigments. Each pigment is assigned a set of absorption coefficients and scattering coefficients, and a granulation constant. We use the KM model to compute the resulting color in RGB space.

User can choose from a range of pigments in the pigment library, and the mixture amount of each pigment. The pigment is then applied to the lit-sphere via a virtual paintbrush, and the flowing and mixing of different pigments is simulated. Flowing effect is isotropic since the paper texture is not simulated in this stage.

After the user finished painting on the lit-sphere, a one-dimensional color-band will be calculated by sampling an arbitrary angle along the sphere's radius. Along with the resulting color in RGB, a granulation variable is also stored in the color-band as the alpha value. The granulation variable defines how visible the granulation effect will be. It is the sum of the granulation constant for each pigment multiplied by the intensity ("thickness") of said pigment in each cell.

3 System Details

The automatic rendering system takes the original 3D geometric models as the input, and apply the watercolor stylization to the 3D scene using vertex and fragment shaders.

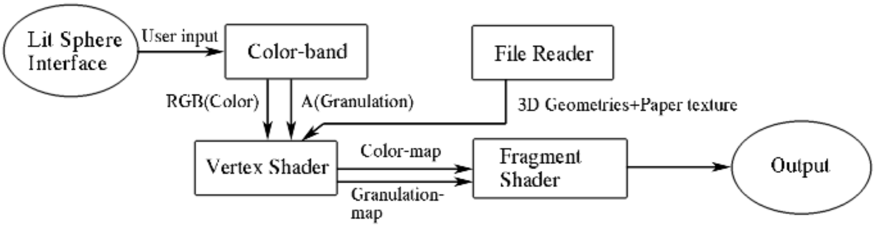


Fig. 1. System Diagram

The shaders we used are written in the NVIDIA Cg language [7]. Two shader scripts are used in generating our watercolor effect. First we take the 3D objects and the color-band to generate a *color-map* and a *granulation-map* using a vertex shader script. Then we process the *color-map* using a fragment shader. The shader takes the *color-map* and a *paper-texture* as the inputs. It combines them with a *Sobel edge map*, which is generated by the shader, to create various watercolor effects.

3.1 Vertex Shader

This part is similar to Lake et al.'s cartoon shading [5]. We return the result of the Phong shading function as a texture coordinate.

`TexCoord0 = Diffuse *` $(N \cdot L)$ `+ Specular *` $(N \cdot H)$

Where N is the normal vector, L is the light vector and H is the halfway vector between the eye vector and light vector. *Diffuse* and *Specular* are the amount of diffuse and specular reflection. In our experiment both are set to 1.

The script is run twice: the first pass takes the RGB element of the color-band as the input, and returns a *color-map*. The second pass takes the alpha value of the color-band (which stores the granulation variable) as the input, and returns a *granulation-map*. Both maps are created using OpenGL render-to-texture capability, and used as the input for the fragment shader.

3.2 Fragment Shader

In this stage, we take the *color-map*, *granulation-map* and the paper texture as input. This shader processes the *color-map* and returns a watercolor styled image, which is our final result. The paper texture is a 2D texture and is treated as a height field. It can be specified by the user or generated using Worley's cellular texturing process [17].

When wet watercolor paint is applied on dry paper, there is a wobbling effect along the edge of the stroke, due to the raggedness of the paper. We simulate this effect first by distorting the *color-map* using the *paper-texture*.

```
colormap_coord = coord+((tex2D(paper_map,coord)-M/2)*D)
```

Where *coord* is the current texture coordinate, M is the mean luminance of the paper texture, D is the amount of distortion. In our experiment M=0.8, D=0.0125. Then we need to refine the image for processing with the Sobel filter. As mentioned in section 2, we use the Sobel filter to create the edge-darkening and granulation effects. Therefore we subtract an amount of *paper-texture* from the original color. The amount of subtraction depends on the *granulation-map*. The higher the granulation value, the more we subtract from the original color. We call the result in this stage a Sobel color-map.

```
Sobel_colormap = org_color -
(tex2D(paper_map, coord)*tex2D(gran_map, coord)*G)
```

Where *org_color* is the original color in *color_map* distorted by the paper texture. G is the weight of granulation effect. In our experiment G=0.5. The Sobel filter detects the discontinuities of the *Sobel color-map*. We intentionally create discontinuities where granulation value is high. When the Sobel filter is applied on the *Sobel color-map*, the resulting edge map will help us create the edge-darkening and granulation effects simultaneously.

After processing the *Sobel color-map* with the Sobel filter (code of which is included in Appendix), finally we put all the results together.

```
out_color = (1-W) + org_color*W - paper_map*P - sobel_edge_map*S
```

The result is a weighted sum of the original color (after distortion), the paper color, and the Sobel edge map, weighted W, P and S respectively. In our experiment, we set $W = 0.6$, $P = 0.1$, $S = 0.3$.

4 Results and Discussion

We used three sample scenes as our examples, shown in Figures 2 and 3. In the "Teapot" scene, we use the KM model to create rich watercolor-style mixing of colors. In the "Bamboos" scene, we can see that even with a simple light-to-dark color-band, our shader can still produce stylish results similar to hand-drawn paintings.

4.1 Performance

We have tested our application on a Pentium 4 3.0GHz machine with an NVIDIA GeForce 5900FX graphics board. All scenes run at about 20 frames-per-second in resolution of 512*512 pixels. The performance depends more on the rendering resolution than the number of polygons, since the fragment shader must perform its operation for every pixel.

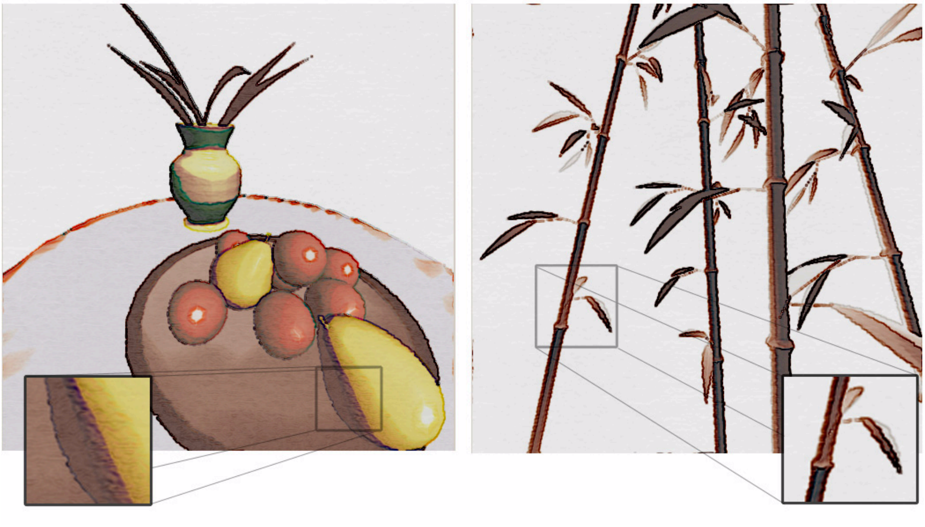


Fig. 2. (Left) Fruit and Vase, 7880 polygons. Notice the different amount of granulation in different paints. (Right) Bamboos, 11080 polygons. Notice how the 2D texture distortion creates the wobbling effect.

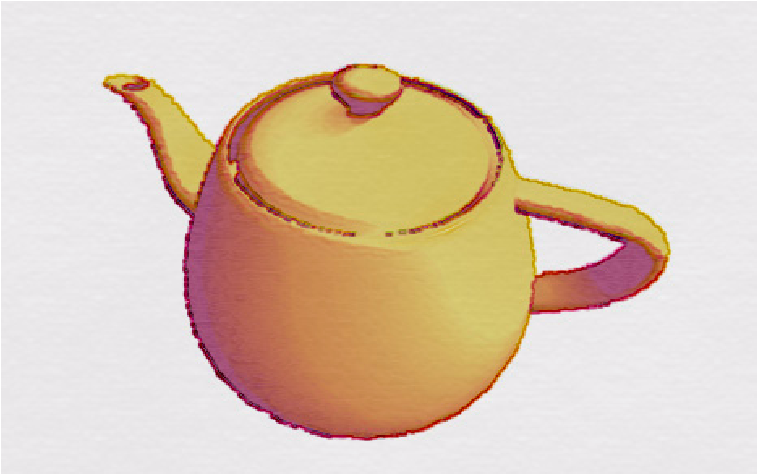


Fig. 3. Teapot, 4096 polygons.

4.2 Conclusions and Future Work

We have introduced a relatively simple approach to render realistic watercolor effects in real-time. We demonstrated by using a combination of color-band shading and the effective use of Sobel filter and 2D scene distortion, we can give the scene a stylish appearance.

For now the flowing effect is simulated only by the color-band and texture distortion. In the future we will try to create a more realistic effect using texture splats [11].

References

1. David Small. “Simulating watercolor by modeling diffusion, pigment, and paper fibers.” In Proceedings of SPIE ’91. February 1991
2. Curtis, C. j., Anderson, S.E., Seims, J. E., Fleischer, K. W. and Salesin, D. H. “Computer-Generated Watercolor.” In Proceedings of SIGGRAPH 87, Computer Graphics Proceedings, Annual Conference Series, edited by Turner Whitted, pp. 421–430, Reading, MA: Addison-Wesley, 1997
3. P. Kubelka. “New contributions to the optics of intensely light-scattering material, part ii: Non-homogeneous layers.” J. Optical Society, 44:330, 1954
4. Peter-Pike J. Sloan, William Martin, Amy Gooch, Bruce Gooch. “The Lit Sphere: A Model for Capturing NPR Shading from Art.” Proceedings of Graphics Interface 2001
5. Lake, A., Marshall, C., Harris, M., and Blackstein, M. “Stylized Rendering techniques for scalable real-time 3D animation.” In NPAR 2000: First InternationalSymposium on Non-Photorealistic Animation and Rendering, edited by Jean-Daniel Fekete and David H. Salesin, pp. 13–20, New York: ACM SIGGRAPH, 2000
6. Freudenberg, B. “A Non-Photorealistic Fragment Shader in OpenGL 2.0.” SIGGRAPH 2002 Talk, San Antonio, July 2002
7. Fernando, R., Kligard, M.J. “The Cg Tutorial: The Definitive Guide to Programmable Real-Time Graphics” Nvidia Corporation, Addison-Wesley, 2003
8. Nienhaus, M., Doellner, J. “Edge-Enhancement - An Algorithm for Real-Time Non-Photorealistic Rendering” WSCG ’03 - The 11-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision’2003, Plzen, Czech Republic, February 3-7, 2003, pp. 346–353
9. Markosian, L., Kowalski, M. A., Trychin, S. J., Bourdev, L. D., Goldstein, D., and Hughes, J. F. “Real-Time nonphotorealistic rendering.” In Proceedings of SIGGRAPH 97, Computer Graphics Proceedings, Annual Conference Series, edited by Turner Whitted, pp. 415–420, Reading, MA: Addison-Weslsy, 1997
10. Willians, L. “Shading in Two Dimensions.” In Graphics Interface ’91, pp. 143–151, San Francisco: Morgan Kaufmann Publishers, 1991
11. Roger A. Crawfis, Nelson Max. “Texture Splats for 3D Scalar and Vector Field Visualization.” IEEE Visualization ’93 Proceedings
12. Gooch, B., Sloan, P. S., Gooch, A., Shirley, P., and Riesenfeld, R. “Interactive Technical Illustration.” ACM Symposium on Interactive 3D Graphics 1999, Atlanta, GA, pp. 31–38, April 1999
13. Raskar, R., and Cohen, M. “Image Precision Silhouette Edges.” ACM Symposium on Interactive 3D Graphics 1999, Atlanta, pp. 135–140, 1999

14. E. Praun, H. Hoppe, M. Webb, and A. Finkelstein, “Realtime hatching,” in Proceedings Computer Graphics (ACM SIGGRAPH) , 2001, pp. 581–586
15. Freudenberg, B., Masuch, M., and Strothotte, T. “Real-Time Halftoning: A Primitive For Non-Photorealistic Shading.” 13 Eurographics Workshop on Rendering. Pisa, Italy, pp. 1–4, June 2002
16. Jason L. Mitchell. “Real-Time 3D Scene Post-Processing”, Game Developers Conference 2003 Talk
17. Steven P. Worley. “A cellular texturing basis function.” In SIGGRAPH ’9 Proceedings, pp. 291–294, 1996
18. Rafael C. Gonzales, Richard E. Woods. “Digital Image Processing.” Addison Wesley, 1993

Appendix: Sobel Filter Implemented by Fragment Shader

```
#define SOBEL_COLOR ORG_COLOR-(tex2D(paper_map, coord)*
    tex2D(gran_map, coord)*G)
float2 coord = tCoords;
coord.x = tCoords.x-offset; coord.y = tCoords.y-offset;
float3 color00 = SOBEL_COLOR;
coord.x = tCoords.x; coord.y = tCoords.y-offset;
float3 color01 = SOBEL_COLOR;
coord.x = tCoords.x+offset; coord.y = tCoords.y-offset;
float3 color02 = SOBEL_COLOR;
coord.x = tCoords.x-offset; coord.y = tCoords.y;
float3 color10 = SOBEL_COLOR;
coord.x = tCoords.x; coord.y = tCoords.y;
float3 color11 = SOBEL_COLOR;
coord.x = tCoords.x+offset; coord.y = tCoords.y;
float3 color12 = SOBEL_COLOR;
coord.x = tCoords.x-offset; coord.y = tCoords.y+offset;
float3 color20 = SOBEL_COLOR;
coord.x = tCoords.x; coord.y = tCoords.y+offset;
float3 color21 = SOBEL_COLOR;
coord.x = tCoords.x+offset; coord.y = tCoords.y+offset;
float3 color22 = SOBEL_COLOR;
float3 Sobel_x = (-1)*color00+(-2)*color01+(-1)*color02
 +(1)*color20 + (2)*color21 + (1)*color22;
float3 Sobel_y = (-1)*color00+(-2)*color10+(-1)*color20
 +(1)*color02 + (2)*color12 + (1)*color22;
float3 Sobel_edge_map = abs(Sobel_x)+abs(Sobel_y);
```

Where offset is the size of a cell in the Sobel filter, which is 1/(color_map.width).

Haptic Interaction in Realistic Multimedia Broadcasting

Jongeun Cha[1], Jeha Ryu[1], Seungjun Kim[2],
Seongeun Eom[2], and Byungha Ahn[2]

[1] Human-Machine-Computer Interface Lab., Dept. of Mechatronics,
Gwangju Institute of Science and Technology,
1 Oryong-dong, Buk-gu, Gwangju 500-712 Republic of Korea,
{gaecha,ryu}@gist.ac.kr,
http://dyconlab.gist.ac.kr

[2] System Integration Lab., Dept. of Mechatronics,
{zizone, seueom, bayhay}@gist.ac.kr,
http://si.gist.ac.kr

Abstract. In this paper, we discuss a haptically enhanced multimedia broadcasting system. Four stages of a proposed system are briefly analyzed: scene capture, haptic editing, data transmission, and display with haptic interaction. In order to show usefulness of the proposed system, some potential scenarios with haptic interaction are listed. These scenarios are classified into passive and active haptic interaction scenarios, which can be fully authored by scenario writers or producers. Finally, in order to show how the haptically enhanced scenario works, a typical example is demonstrated to explain specifically for a home shopping setting.

1 Introduction

Rapid development of computing and telecommunication technology such as enhanced CPU speed and power, low cost memory, and ultra fast communication network has led to digital multimedia age, where viewers can be immersed in some 3D visual and audio contents. Moreover, viewers can interact even haptically with the multimedia contents beyond passive watching and listening. Traditionally, these interaction and immersion are possible only with fully virtual worlds (VR), in which the world is filled with synthesized objects. In addition, only one viewer or small number of viewers can share the virtual contents at the same time [1].

In the area of broadcasting system, new technology is being developed and is available in terms of digital multimedia broadcasting through the air or through the Internet. Main multimedia contents are, however, limited to 2D video and sound so that feeling of full immersion is still far from the reality. Interactivity is also being pursued in these days in a very simple form such as selection and retrieval of 2D AV contents. In the report of the Media Interaction group of Philips Research in Eindhoven, recent efforts for the interactive television are

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 482–490, 2004.

well summarized from its concepts and history to storytelling application [2]. If the broadcasting network in the future can be completely integrated with communication network like Internet, useful techniques in the network services such as chatting program, server-client system, web casting, device communication can be technically available for the broadcasting system. All these, in some scenarios, may make us enjoy attractive bi-directional services by immersing dynamically into the broadcasting productions that may include sense of touch if viewers want to fully interact with more realistic multimedia contents.

With ATTEST project, which started in March 2002, the development of the first commercially feasible European 3D-TV broadcast system has been in progress. In [3], 3D-video chain of ATTEST including 3D content creation, encoding, transmission, and display stages is described. They have been trying to use head tracking to drive the display optics and develop two 3D displays, one for a single viewer and one for multiple viewers. At the same time, O'Modhrain and Oakley [4,5] discussed the potential role that haptic or touch feedback might play in supporting a greater sense of immersion in broadcast content. Presenting Touch TV, they showed two potential program scenarios: the creation of authored haptic effects for children's cartoon and the automatic capture of motion data to be streamed and displayed in the context of a live sports broadcast.

Unlike the simple addition of touch-enhanced contents to the broadcast media in some scenarios, in this paper, we are investigating more comprehensive realistic multimedia broadcasting system that can include haptic interaction in addition to 3D audio-visual contents. More specifically, firstly, we present a top-level view of creating, editing, transmitting, and displaying with viewer interaction fully immersive multimedia contents in a broadcasting system through the Internet. This is a new attempt beyond a multicasting system that utilizes small number of shared computational platforms servicing for ten to hundred viewers only. We describe each stage of the proposed system in terms of data type and generic processing algorithm, etc. Secondly, in order to present usefulness of the proposed system, we have listed some possible scenarios with haptic interaction. Producers who may be science teachers/professors, educationers, geographers, artists, etc can develop specific sense-of-touch-added scenario. Role of engineers is to provide these producers with content creation tools such as sensors embodied in a real or virtual object, multimedia authoring tools, interaction techniques and devices with haptic sensation. Then viewers can enjoy the immersive interaction dynamically to get their indirect experience as well as additive information. Finally, we present a typical application example to show how the haptically enhanced scenario works. This simple demo system utilizes Augmented Reality (AR) techniques, which show excellence in synthesizing seamless videos in real-time, multimedia streaming technology, and a 6 degree-of-freedom haptic device for a homeshopping setting.

2 Haptically Enhanced Multimedia Broadcasting System

From the high-level view, a general broadcasting system can be divided into four parts: capture, edit, transmission, and viewing. A producer captures a scene

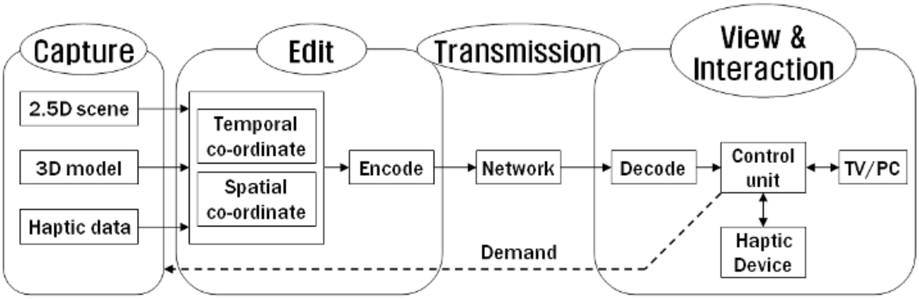


Fig. 1. Haptically enhanced multimedia broadcasting chain

with a camera and edits that scene to make a broadcasting content by cutting, sticking, synthesizing computer graphics and audio contents, and so on. Then the authored program is transmitted to the viewer via the airwave, satellite, cable, or Internet. The viewer passively watches the program in front of the TV with a remote control in his hand for simple interaction, for example, changing the channels.

Figure 1 shows the proposed haptically enhanced multimedia broadcasting system. In this figure, contrary to the traditional broadcasting system, the video media, which is the sequence of the 2.5D scenes plus virtually synthesized computer graphics models, has the depth information in addition to the 2D image. It depicts the geometry of the captured scene in terms of 3-dimensional coordinate from the camera view, not the arbitrary view. 3D computer graphics models can be easily registered with the 2.5D scene in a 3-dimensional space. In addition to the video media, haptic data, the authored, recorded, or physically sensed media representing the kinesthetic and tactile information, e.g. material property data at each pixel for texture tactile feeling or object weight data associated with an object-of-interest for force sensation, is combined to give the viewer haptic effects. The edited hyper media, which can be defined as 3-dimensional audio-video media synchronized with the haptic data, is transmitted through encoding and decoding operations to the viewer site via the Internet. The control unit receiving the media renders the stereo images and 3D sound to the display device by processing the 3-dimensional audio-video media and controls the haptic device to give a haptic interaction to the viewer. In this way, the viewer can actively interact with the 3-dimensional hyper media as well as can feel the haptic effects. Besides viewers can also demand additional data via the bi-directional Internet channel.

In the capture stage, all scenes are captured with the depth information. A depth camera like Zcam ^{TM} [7] yields conventional 2D-video accompanied with depth-per-pixel via a direct depth-sensing process. Multiple 2D cameras can generate the same information even though the overall processing is more complicated than the direct depth sensing [6]. A 3D scanning and subsequent reconstruction method or modeling with a CAD program may capture virtually synthesized 3D models. These two kinds of data are composed of the 3-dimensional video media. 3D audio sound may also be produced by new technology. In the meantime, some authored haptically enhanced data such as vibration of bee wings in the kid's animation, kicking force of a soccer ball, which may be obtained in reality by a real accelerometer embedded inside a ball, etc may also be captured along with the audio-visual data [5]. Notice that in order to synchronize the haptic data with the audio-visual data, haptically related data such as position, velocity, and acceleration data of the corresponding object of interest should also be recorded.

In the edit stage, the captured data are coordinated temporally and spatially to a program. Since the handled data is basically 3-dimensional, the composition operation needs to be managed in the 3-dimensional space. The 3D model is synthesized to the captured 2.5D scene by the z-keying or the Augmented Reality technique. The z-keying is the process to put the 3D model anywhere in the captured scene by giving depth. The Augmented Reality technique is a process to place the 3D model on a specific position, utilizing a specific feature in the captured scene. In addition, the captured haptic data is synchronized with the video-audio data utilizing the haptically related data in this stage.

In the transmission stage, the finished program is transferred to viewers via Internet. Because the communication channel is Internet, many useful bi-directional interactions between viewers and broadcasters may be possible. Viewers can demand some data from the broadcaster especially in a live broadcasting system.

In view and interaction stage, the control unit processes the video media and the haptic data to produce a stereoscopic scene to the display device and to control haptic device by using haptic rendering algorithm. In haptic interaction stage, viewers may feel the transmitted haptic effects that are synchronized with some specific scenario passively by putting their hands on the vibrotactile display device. Or they can actively touch, explore, and manipulate some transmitted 3D CG objects or background 2.5D scene according to the preplanned path or to the viewers' will.

3 Potential Haptic Interaction Application Scenarios

A distinctive characteristic of the broadcasting program is that it is captured and edited nonlinearly but broadcasted linearly in time domain. The program, as any other television show, has to be waited for, i.e., has a fixed position in the daily schedule. Furthermore, it is not possible to go back, or start the program over. Therefore, in this paper, the haptic interaction occurs in the viewer site only and does not change the path of the program story. But viewer can explore the program or manipulate some digital objects by active touch or by the passive haptic effect that is authored and provided. In this section, we list some potential application scenarios that take advantage of the haptic interaction.

Potential haptic interaction scenarios may be classified into passive or active interaction: Passive haptic interaction scenario just records some haptically related data when capturing audio-visual scene including object-of-interest and sends them to the viewer with interaction time indicator (e.g. caption on the screen). Then, the haptically-related data controls a haptic device worn on the viewer's hand or arm. In this interaction, therefore, viewers are only passive. Active haptic interaction scenario captures 2.5D audio-visual scene only or together with full 3D virtual objects of interest that are either independent of the 2.5D scene or dependent on it as is the case of Augmented Reality. Then, these data are transmitted to the viewer's control box, where 2.5D scene, virtual objects, and haptic device are synchronized temporally and spatially. After this, viewers can interact actively with the scene or object-of-interest, i. e., can touch or push buttons. Collision detection and response calculations are all done in the viewer's control box in this case. In the following potential scenarios, scenario 1 is passive and fully authored by producer. This passive interaction may be lively broadcasted, e.g. live soccer game. Meanwhile, scenarios 2, 3 & 4 are active and viewers take time to explore or manipulate the virtual object-of-interest. This active mode may be live too.

3.1 Scenario 1: Feeling Haptic Data

In teaching programs of some manipulation techniques, a producer may want the viewer to follow the instructor's movement because it is very useful to learn expert's manipulation technique by viewing his actions as well as by tracking. For example, the expert in pen writing is showing how to write a pretty hand. He asks the viewer to grip the pen-like haptic device and starts to write a character. The viewer is completely guided to move the pen following the expert's writing. In this case, the producer captures a handwriting expert visually as well as haptically by recording the hand poses in real time or off-line. The captured scene and the recorded pose data are edited synchronously to an educational content. The control unit in the viewer's site displays the expert's handwriting and controls a haptic device to follow the recorded hand pose with the viewer wearing the pen-type haptic device. Note that this scenario is record-and-play type haptic interaction. This scenario does not record haptic data but record pose (position and orientation) data in the capture stage. In the interaction display stage, recorded pose data will drive the haptic device with force generation so that viewers feel touch sensation.

3.2 Scenario 2: Touching 2.5D Scene

While viewing a TV program, a viewer sometimes may want to touch a weird shaped thing or an actor's face to acquire the shape or the skin feeling. For example, in a drama, lovers are looking into the eyes of each other and going closer. A viewer may want to touch one of the acting lovers on the face. His face is slowly closed up with a camera and the 'haptic interaction' caption shows up on the corner of the screen. The viewer then touches the actor's face by a haptic device worn on the hands. This touching interaction may be possible if the video media has 2.5 or 3-dimensional information. For this scenario, a program is captured as the 2.5D scene that contains object-of-interest to touch. Since the viewer interacts with the object by the physical force contact, in the captured scene the object-of-interest should be static or moving slowly. The producer overlays a haptic interaction caption indicating when to touch. When a viewer wants to touch, the control box will perform collision detection and force computation to drive a haptic device worn on viewer's hands.

3.3 Scenario 3: Touching and Manipulating 3D Models

Sometime, a producer may want to let the viewer feel something in a program as well as see it to give deeper understanding. For example, in an education channel for the science of dynamics an instructor is explaining the force equation for a spring. After teaching the theory, he uses augmented reality technology to arrange few virtual springs on a real experiment desk and asks the viewer to push the spring and feel the spring force with the haptic interaction caption. Then he repeats the experiment changing the number and the arrangement of the springs. In this scenario, the viewer can have deeper knowledge of the spring properties by handling it directly as well as being taught the theory. For this, a program is captured as 2.5D scene including the clue, like a feature, for composing the realistic looking 3D model using Augmented Reality technique. In the program, the MC puts the features and makes the experimental environment seeing synthesized video in real time. After capturing the program, the producer augments the 3D model, such as a virtual spring, exactly and stably and attaches the haptic interaction caption. A viewer follows the program and interact haptically. This scenario is different from the previous scenario in touching and manipulating 3D virtual objects instead of 2.5D object that is captured from the real scene.

3.4 Scenario 4: Touching and Manipulating 3D Models on Demand

Sometimes, a viewer may want to buy a product-of-interest while watching a TV program. In this case, providing him with haptic sensation in addition to the audio-visual information of the product can help his purchasing decision. For example, while viewing a drama, a viewer may get interested in a camera that the actor is using. He picks up a remote controller and pushes the menu button to get the camera product information. After looking into the specification, he may get into the haptic mode to try to touch the camera. He, then, can manipulate some buttons, or touch the surface feature, or feel the inertia of the downloaded virtual camera. For this scenario, a producer makes an advertisement's content that includes the product information and 3D virtual model of the product and saves them in the server. The viewer demands the additional information for the product, downloads it, and examines the product carefully for purchasing by exploring and manipulating it.

4 Demonstration Example

This section explains a demonstration example of an active haptic exploration/ manipulation in a home shopping setting. To explore how the scenario works,

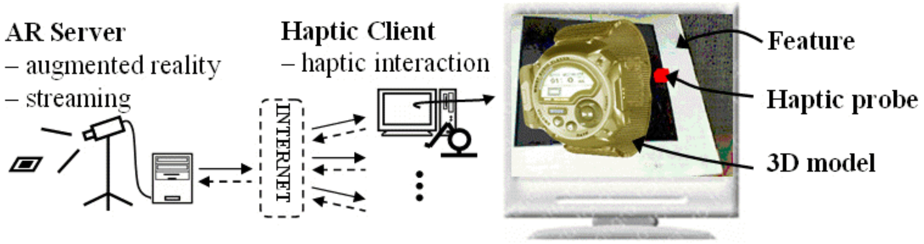


Fig. 2. Demonstration system overview

we have implemented a simple broadcasting system, which is capable of delivering video media on the Internet and giving haptic interaction. The demo system is constructed basically following the stages of the broadcasting chain in Fig. 1, with two major subsystems: AR (Augmented Reality) server and Haptic client. As shown in Fig. 2, the AR server consists of a typical AR system and a streaming server. This server makes it possible to create a broadcasting content and stream it via Internet. Therefore, Capture, Edit and Transmission stage are all performed in this server. The View & Interaction stages are implemented in the Haptic client system. It receives the content and realizes viewer's interaction with a connected haptic device. We have used Augmented Reality techniques based on ARToolKit[9], multimedia streaming technique, and a 6-dof haptic system (PHANTOM[10]) in the demo system. ARToolKit is a software library that can be used to calculate camera position and orientation relative to physical markers in real time. The SensAble Technologies PHANTOM makes it possible for users to touch and manipulate virtual objects through the help of the GHOST SDK (General Haptic Open Software Toolkit) that is a powerful, C++ software tool kit that eases the task of developing touch-enabled applications. Firstly, we explain how the system works technically and then show how the home shopping application scenario can be realized. The AR system captures the real environment scene that contains a known marker and obtains the position and orientation of the marker relative to the camera. Then, the streaming server packages the captured scene and the marker's location and transfers them via Internet. The 3D model data is transferred in advance through the other channel. At viewer's end, the haptic client system receives the transferred data and augments the 3D model to the captured scene at the marker's location. The haptic probe, that corresponds to the handle of a haptic device grasped by a viewer, is graphically overlaid to the augmented scene relative to the camera reference. The viewer is able to interact with the 3D model by moving the handle as watching the scene.

In the example application scenario, we have considered a situation that a shopping host tries to advertise a product: Wrist-held MP3 player. She wants to explain functions and features of the product using visual and haptic sensation. She explains the product by rotating it and making viewers touch surfaces or push buttons. When the host puts a marker-based feature in the camera range, viewers in the haptic client site can watch a scene augmented with the MP3 player 3D model. Then, they grasp the handle of haptic device to interact with it. Watching the haptic probe navigating in the scene according to the each viewer's intention, they can actively explore the outlines of the product and push a functional button to know how it works.

We have implemented the first demo example based on the broadcasting chain, which is haptically enhanced. It makes us fully immersed into the broadcasted world and provides much interest in experiencing the broadcasted contents.

5 Future Work

As discussed in the previous sections, haptic interaction in a broadcasting system requires new data format and processing technique in each stage. For example, passive haptic-related data must be prepared in the capture/edit stage and be transmitted along with the audio-visual data plus some camera-related data. In addition, conventional haptic rendering algorithms were mainly developed focusing on the interaction between 3D models in virtual environments. Moreover, the process to get the 3D model by scanning the real object or modelling with a CAD program is time-consuming. The advent of the Zcam^{TM}[7], however, makes it relatively easy to get the 2.5D model of the real scene because the process is just capturing not modelling. In a scenario where complete 3D model is not needed e.g. scenario exploring only the visible part inside a view fulcrum, the scene data will be 2.5D and a novel haptic rendering algorithm (collision detection and response calculation) for the 2.5D scene is needed.

In this paper, we consider only the interaction between viewers and the broadcasting multimedia. The noticeable feature of the interaction in the haptics is the social presence, the feeling of being socially present with another person at a remote location[8]. In a live broadcasting scenario, an audience may take part in a program and communicate with other people viewing the same program by haptic interaction. Since the communication channel of the broadcasting via the Internet is bi-directional, it seems to be possible. One can pursue to establish the system and the scenarios for the social presence in the future.

6 Conclusions

In this paper, we discussed a top-level structure and brief data structure and processing algorithm of future realistic multimedia broadcasting system that may include sense-of-touch. Also, some potential scenarios taking advantages of haptic interaction were listed in a realistic broadcasting in which the video media is 3-dimensional. Finally, an application example demo system is presented. Addition of the haptic interaction to the conventional audio-visual contents will improve the immersion of the viewers together with rich contents. Moreover, full engagement to the realistic multimedia by haptic interaction can enhance amusement as well.

Acknowledgements. 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] Grigore C. B., Philippe C.: Virtual Reality Technology. Second Edition, by John Wiley & Sons, (2003)

[2] Bukowska, Magdalena: Winky Dink half a century later : interaction with broadcast con-tent : concept development based on an interactive storytelling application for children. (2001)

[3] Andre R., Marc O. B., Christoph F., Wijnand I., Marc P., Luc V. G., Eyal O., Ian S., Philip S.: Advanced Three-dimensional Television System Technologies. ATTEST Publication, Padova, Italy

[4] O'Modhrain, S., Oakley, I.: Haptic Interfaces for Virtual Environment and Teleoperator Systems. HAPTICS '04. Proceedings. 12th International Symposium on. (2004) 293–294

[5] O'Modhrain S., Oakley I.: Touch TV: Adding Feeling to Broadcast Media. in proceedings of the European Conference on Interactive Television: from Viewers to Actors, Brighton, UK, (2003) 41–47

[6] Grau O., Price M., Thomas G. A.: Use of 3-D Techniques for Virtual Production. BBC R&D White Paper, WHP 033, (2002)

[7] 3DV Systems. http://www.3dvsystems.com

[8] E. Sallnas, K. Rassmus-Grohn and C Sjostrom: Supporting Presence in Collaborative Environments by Haptic Force Feedback, ACM Transactions on CHI 7(4), ACM Press, 2000, 461–476

[9] ARToolKit computer vision software: http://www.hitl.washington.edu/artoolkit

[10] PHANTOM, SensAable Technologies, http://www.sensable.com/

Object-Based Stereoscopic Conversion of MPEG-4 Encoded Data

Manbae Kim, Sanghoon Park, and Youngran Cho

Kangwon National University
Department of Computer, Information, and Telecommunication
192-1 Hoja2-dong, Chunchon 200-701, Republic of Korea
manbae@kangwon.ac.kr

Abstract. Stereoscopic conversion of two-dimensional (2-D) video is considered in object-based approach that independently processes each video object. Our works extend the previous frame-based stereoscopic conversion of MPEG-1 and 2 to MPEG-4. In MPEG-4, each image is composed of a background object and primary object(s). In the first step, a camera motion type is determined for generating a stereoscopic background image. For this, motion vectors of a background object are utilized. The generation of a stereoscopic background object makes use of a current image and a previous image. As well, The stereoscopic primary object uses a current image and its horizontally-shifted version to avoid the possible vertical parallax that might happen. In the second step, the two stereoscopic objects are combined to generate a stereoscopic image. As verified in experiments performed on two MPEG-4 sequences, the object-based stereoscopic conversion can cope with a vertical parallax that has been a difficult problem to deal with in the frame-based approach.

1 Introduction

Stereoscopic video enables the three-dimensional (3-D) perception by producing the binocular disparity existing between left and right images. In general, a stereoscopic camera with two sensors is required for a stereoscopic video. In contrast, stereoscopic conversion directly converts 2-D video to 3-D stereoscopic video. The principle of the stereoscopic conversion stems from the psychophysics theory of Ross [1,2], in which stereoscopic perception can be generated by combining the current and delayed images displayed appropriately to both human eyes. However, a main constraint is that image motions of camera, object or both need to be horizontal. In other words, vertical motions could cause visual discomfort [7] and thus pose a difficulty in producing a stereoscopic image.

There are reported many research works on the stereoscopic conversion of NTSC signal [3,4] as well as MPEG-1 and 2 [5] based upon the Ross phenomenon. In those methods, dealing with the vertical motions needs complex algorithms, thereby requiring high computational complexities. In MPEG-1 and 2, video data are encoded in the frame-based manner, thereby posing a difficulty in processing

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 491–498, 2004.

the vertical motions. On the other hand, the stated constraint can be overcome in MPEG-4, where an image sequence is encoded in the object-based manner.

In MPEG-4, a frame is composed of video object planes (VOPs) and each VOP is independently encoded [6]. Luma and chroma data are encoded using shape as well as texture information. In summary, our proposed method extends such earlier works on the frame-based stereoscopic conversion of MPEG-1 and 2 to support the object-based stereoscopic conversion of MPEG-4 that this paper is mostly concerned with. Each frame is assumed to have a background VOP. It is also assumed that a single primary VOP exists in the image. The two VOPs are independently processed to generate stereoscopic VOPs. Then, they are combined together and a stereoscopic image is produced. Furthermore, the stereoscopic conversion of multiple primary VOPs can be easily handled without any major revision.

The following section describes the principle underlying the stereoscopic conversion of 2-D video. Section 3 presents our object-based stereoscopic conversion. Some of experimental results are reported in Section 4 followed by the conclusion of Section 5.

2 Principle of Stereoscopic Conversion

The stereoscopic conversion of 2-D image with a horizontal motion is shown in Fig. 1. Suppose that the image sequence is $\{\cdots, I_{K-3}, I_{K-2}, I_{K-1}, I_K, \cdots\}$ and I_K is the current frame. Then, a stereoscopic image consists of I_K and one of the previous frames, I_{K-i} $(i \geq 1)$. If the current and previous images are appropriately presented to both human eyes as in Table 1, then the user feels the 3-D stereoscopic perception [5]. For example, in the case of a right motion of a camera and no object motion, previous and current images need to be displayed to left and right eyes, respectively. As mentioned in the previous section, the usage of previous images based upon the Ross phenomenon is applied only to the horizontal motion so that it is difficult to apply conventional conversion methods to other types of motion such as non-horizontal, zooming, fast motion,

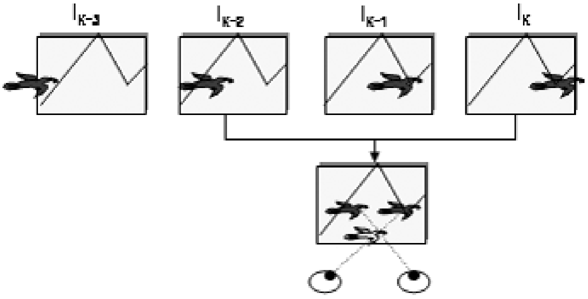


Fig. 1. shows how a stereoscopic image is produced from images with a horizontal motion

Table 1. The selection of left and right images according to camera and object motions

Camera Motion	Object Motion	Left Image	Rigth Image
Right	None	Previous	Current
Left	None	Current	Previous
None	Right	Current	Previous
None	Left	Previous	Current

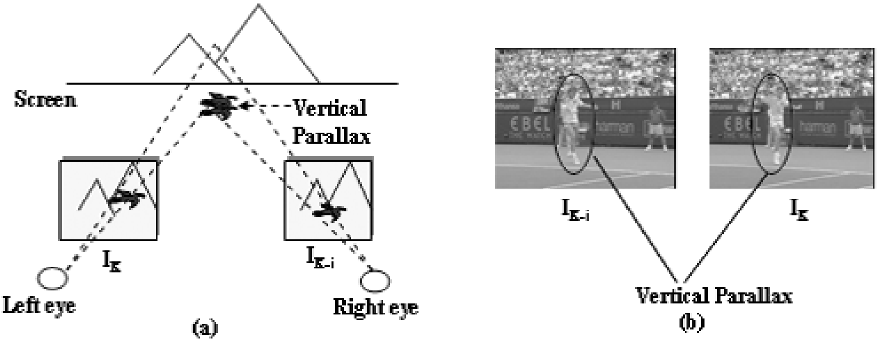


Fig. 2. shows examples of vertical parallax. In (a), a bird is moving in non-horizontal direction, causing the vertical disparity in the left and right images. As well, it is observed that some regions of a tennis player generate a vertical disparity due to a random movement

and so forth. Fig. 2 illustrates examples of non-horizontal motions. In (a), a camera is fixed and an object (bird) is moving in the non-horizontal direction. The background has zero parallax or disparity due to no movement. On the contrary, a vertical parallax appears due to the non-horizontal movement of the object. (b) shows an example of a non-rigid object (a tennis player). The vertical parallax between the two successive images is observed due to the non-rigidity characteristics. Most of the previous works convert PAL/NTSC signal and MPEG-1 and 2 data to 3-D stereoscopic video. Since they are processed in the frame-based manner, it is difficult to process non-horizontal motion images. On the other hand, MPEG-4 encodes images in the object-based manner [6]. Our proposed object-based stereoscopic conversion of MPEG-4 encoded data can be easily carried out even for non-horizontal motions.

3 Proposed Method

This section presents a methodology underlying a stereoscopic conversion as shown in Fig. 3. The input is an MPEG-4 video bitstream being composed of two video object planes (VOPs): a background object, $\mathrm{VOP}_{\mathrm{BO}}$ and a primary object, $\mathrm{VOP}_{\mathrm{PO}}$. They are separated into encoded $\mathrm{VOP}_{\mathrm{PO}}$ and $\mathrm{VOP}_{\mathrm{BO}}$ by a demultiplexer. Each decoded VOP is processed by the stereoscopic image gener-

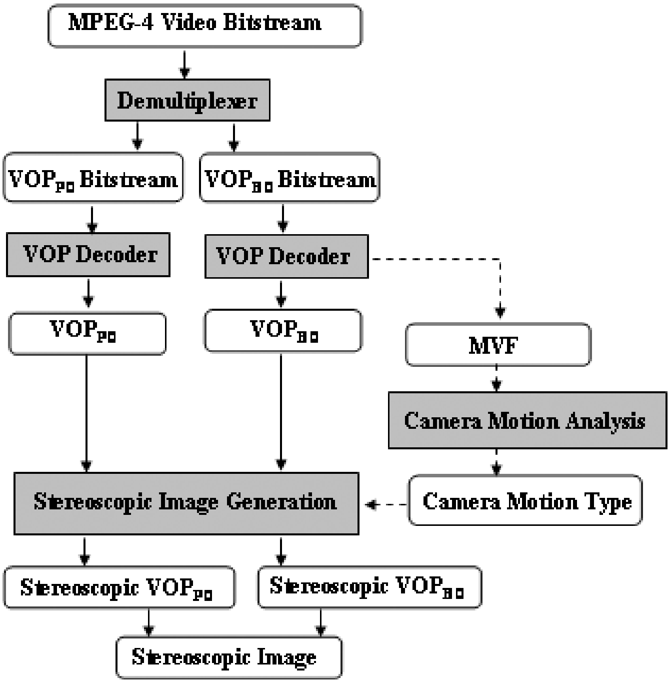


Fig. 3. An overview of an object-based stereoscopic image generation

ation. The camera motion analysis processes the motion vector field (MVF) of $\mathrm{VOP_{BO}}$ and determines a camera motion type among left motion, right motion, and static. The MVF is the set of all motion vectors (MVs) of 16×16 macoblocks. $\mathrm{VOP_{PO}}$ does not affect the decision of the camera motion type. A basic idea of the object-based stereoscopic image generation is to perform separate operations on two VOPs and produce stereoscopic $\mathrm{VOP_{PO}}$ and $\mathrm{VOP_{BO}}$. Finally we combine them to generate a stereoscopic image. For determining the types of a camera motion, we compute the camera displacement, $(\mathrm{dx_{cam}}, \mathrm{dy_{cam}})$ by independently averaging x and y components of the MVs of $\mathrm{VOP_{BO}}$. Then, the camera motion type is determined by a sign value of $\mathrm{dx_{cam}}$. Plus and minus signs indicate the left and right camera motions, respectively. On the other hand, $\mathrm{dy_{cam}}$ is used for shifting either a left or a right image vertically in order to reduce a vertical disparity or parallax between the two images. As mentioned before, the vertical parallax causes visual discomfort and thus needs to be reduced as much as possible.

A stereoscopic primary object is composed of a primary object of a current image and its horizontally-shifted object. A shift value, $\mathrm{S_{PO}}$ ranges at $[0, \mathrm{T_P}]$, where $\mathrm{T_P}$ is defined as a maximum human perception threshold [7]. $\mathrm{S_{PO}}$ is proportional to 3-D depth and needs to be less than $\mathrm{T_P}$. Figure 4 illustrates how a stereoscopic primary object is generated according to a camera motion type. It is assumed that a negative parallax applies to primary objects for better

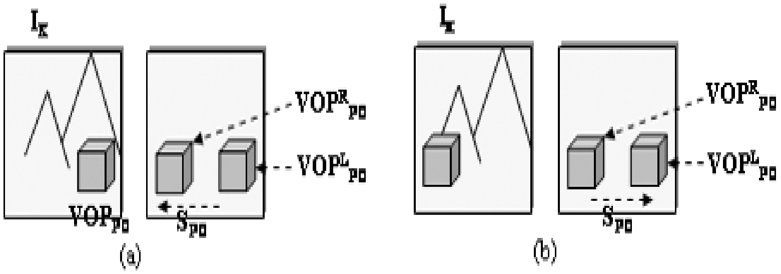


Fig. 4. Stereoscopic processing of $\mathrm{VOP_{PO}}$. The camera motion types are (a) right motion and static, and (b) left camera motion

3-D depth. In (a) of a right motion of a camera and a static camera, $\mathrm{VOP^L_{PO}}$ indicates $\mathrm{VOP_{PO}}$ of a left image. The current image I_K becomes the left image. The object is shifted to the left direction by $\mathrm{S_{PO}}$ and is denoted by $\mathrm{VOP^R_{PO}}$ in the right image. For (b) of a left motion of a camera, a primary object of the current image is located in the right image, and its shifted object is in the left image. left and right images.

Unlike the primary object, processing a background object, $\mathrm{VOP_{BO}}$ makes use of a background object (image) of current and previous images. To do this, we need to determine one of previous images, which is chosen by a delay factor, $\mathrm{f_D}$ being computed by

$$\mathrm{f_D} = \mathrm{ROUND}[\frac{\mathrm{T_D}}{|\mathrm{dx_{cam}}|}] \tag{1}$$

where $\mathrm{T_D}$ a threshold for a maximum displacement defined by a user and $\mathrm{dx_{cam}}$ is a horizontal camera displacement. A previous image chosen is I_{K-f_D}

From VOPs in the current and previous images, the determination of left and right $\mathrm{VOP_{BO}}$s depends upon a camera motion type. For a static type, the current VOP is applied to both VOPs. The current and previous VOPs become left and right ones for a left motion, respectively. On the contrary, the order is reversed for a right motion. Furthermore, if any vertical motion exists, we need to shift $\mathrm{VOP_{BO}}$ by $\mathrm{S_{BO}}$ being computed by

$$\mathrm{S_{BO}} = \mathrm{dy_{cam}} \cdot \mathrm{f_D} \tag{2}$$

where $\mathrm{f_D}$ scales the interval between a current image and its associated previous image.

4 Experiments

This section contains the results of some experiments performed for validating our proposed method as well as examining 3-D stereoscopic perception. The

Table 2. Performacen results of camera motion analysis

Test Data	No. of Frames	N_R	N_S	N_L	Accuracy (%)
Stefan	100	49 (46)	32 (32)	19 (22)	94
Coastguard	100	57 (63)	18 (22)	15 (15)	91

experimental results are presented using two MPEG-4 test sequences: Stefan and Coastguard. The size of the image is 352 × 288 and each image has two VOPs. The number of test images is 100 for each sequence. The accuracy rate of the camera motion analysis is shown in Table 2. N_R, N_S, and N_L are the number of frames for right motion, static, and left motion of the camera, respectively, obtained from our camera motion analysis. $(\cdot)$ indicates the number of frames with correct motions. The accuracy ratio obtained from the two image sequences is approximately 93%.

Figure 5 shows the results of the stereoscopic conversion of *Stefan* sequence. The large movement of a non-rigid object (a tennis player) is observed in the sequence. In (a), the four images are a current (left) image, a primary object, a background object, and MVF for the current image. (b) shows a right image, its primary object, its background object, and an interlaced stereoscopic image of the left and right images. The camera motion analysis of the MVF results in the horizontal displacement of -12, so that the camera motion type is a right motion.

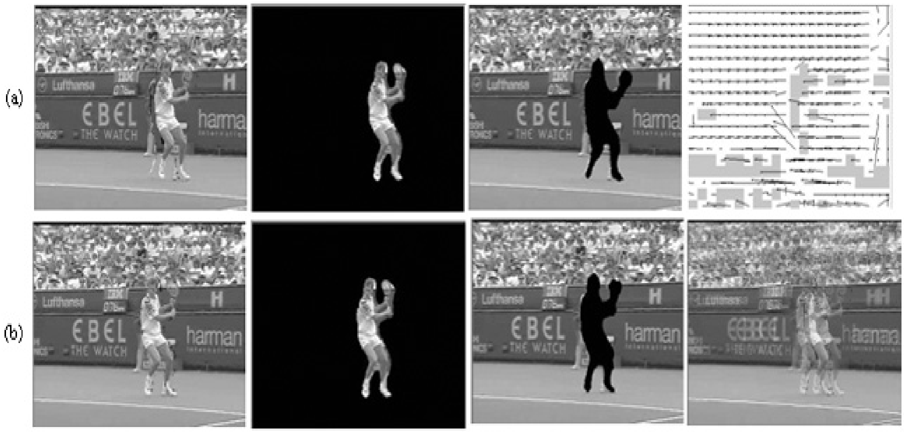


Fig. 5. Test results of Stefan sequence

The result of the Coastguard sequence is shown in Fig. 6 for two different images. The vertical movement of the camera is zero indicating static camera. Therefore, the same background object of a current image is used for left and right images. For (a) and (b), the four images are a left image, a right image, MVF, and an interlaced image.

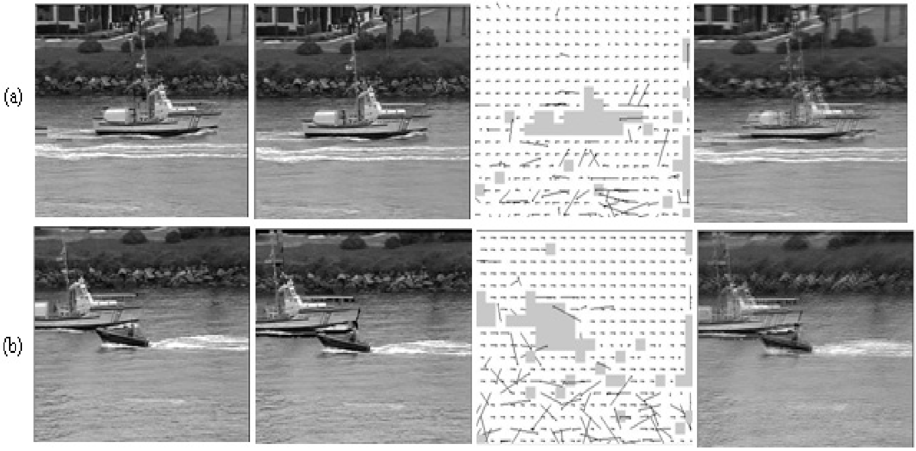


Fig. 6. Test results of Coastguard sequence

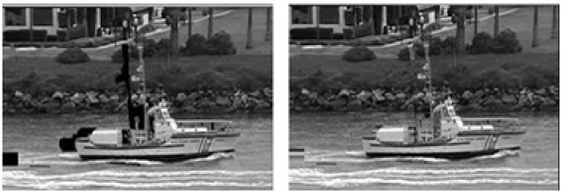


Fig. 7. UR processing. (a) Before filling UR, (b) After filling UR

Uncovered region (UR) would appear at either the left or right image due to shifting of a primary object. Such regions need to be filled with appropriate colors, posing a difficult problem to solve for. We filled the UR with partial areas of a previous image. The details will not be presented due to the space limit. An example of filling the UR is shown in Fig. 7. In (a), some UR regions are observed (black regions) and the image after filling is shown in (b).

5 Conclusion

In this paper, we have presented a stereoscopic conversion of object-based encoded data. This work extends previous works on the frame-based stereoscopic conversion of MPEG-1 and 2 to support the object-based processing of MPEG-4. To do this, each VOP is independently converted into a stereoscopic VOP that are combined to form a stereoscopic image. The main advantage of the object-based conversion is that it can easily deal with non-horizontal motions that have been a difficult problem in the frame-based conversion schemes.

The experimental results validate the feasibility of our proposed method as well as the easy implementation compared with other conventional methods.

Further investigation with two or more VOPs would be necessary. In this case, one of major difficulties is to derive a relative 3-D location of each object that is important to perception of 3-D depth.

Acknowledgments. This work was supported by the Ministry of Information and Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.

References

1. Ross. J. and Hogben, J. H., "Short-term memory in stereopsis," In Vision Research, Vol. 14, pp. 1195–1201, 1974
2. Burr, D. C. and Ross, J., "How does binocular delay give information about depth?," In Vision Research, Vol. 19, pp. 523–532, 1979
3. T. Okino and et al., "New television with 2-D/3-D image conversion technologies," SPIE Vol. 2653, Photonic West, 1995
4. B. J. Garcia, "Approaches to stereoscopic video based on spatial-temporal interpolation," SPIE Vol. 2635, Photonic West, 1990
5. Man Bae Kim and Sang Hun Lee, "A new method for the conversion of MPEG encoded data into stereoscopic video", J. of the Society for 3-D Broadcasting and Imaging, Vol 1. No. 1, pp. 48–59, June 2000
6. P. Fernando (edited), The MPEG-4 Book, Pearson Education Inc., 2002
7. D. F. McAllister (edited), Stereo computer graphics and other true 3-D technologies, Princeton, NJ: Princeton University Press, 1993

Shared Annotation Database for Networked Wearable Augmented Reality System

Koji Makita, Masayuki Kanbara, and Naokazu Yokoya

Graduate School of Information Science,
Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
{koji-ma, kanbara, yokoya}@is.naist.jp

Abstract. This paper describes a database of annotation information for augmented reality (AR) on wearable computers. With the advance of computers, AR systems using wearable computers have received a great deal of attention. To overlay annotations on the real scene image, a user's computer needs to hold annotation information. The purpose of this paper is to construct a networked database system of annotation information for wearable AR systems. The proposed system provides users with annotation information from a server via a wireless network so that the wearable computers do not need to hold it in advance and information providers can easily update and add the database with a web browser. In experiments, the user's position-based annotations have been proven to be shown to the user effectively.

1 Introduction

Since computers have made a remarkable progress in resent years, a wearable computer can be realized [1]. At the same time, the augmented reality (AR) technique which merges the real and virtual worlds has received a great deal of attention as a new method for displaying location-based information in the real world [2–4]. Therefore, AR systems using wearable computers will open up a new vista to the next generation wearable computing [5,6]. Figure 1 shows an example of annotation-overlay using a wearable AR system. Since the wearable AR system can intuitively display information to user on the real scene as shown in Figure 1, it can be applied to a number of different fields [5,7–12]. To realize a wearable AR system, the position and orientation of user's viewpoint and annotation information are needed. The position and orientation of user's viewpoint are needed for acquiring the relationship between the real and virtual coordinate systems. Many researchers have proposed a number of different methods for measurement of the position and orientation of user's viewpoint with some kinds of sensors [6,7,13–15]. To overlay annotations on the real scene image, a user's computer needs to hold user's location-based information. Up to this time, since a database of annotation information is usually held in the wearable computer in advance, it is difficult for the database of annotation information to be easily updated or added by information providers (including normal PC users and wearable PC users).

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 499–507, 2004.

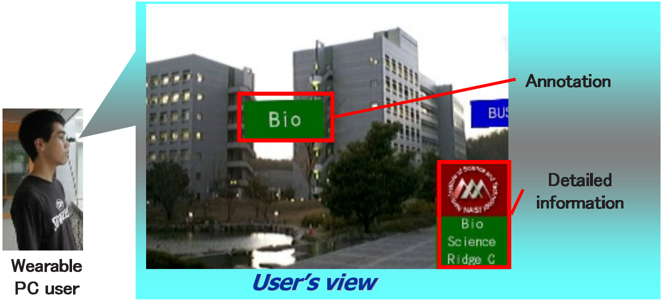


Fig. 1. An example of annotating a real scene.

The purpose of the present work is to construct a shared database system of annotation information for wearable AR systems. To realize the system, we install a database server which can be accessed with a wireless network. The database is shared by multiple users of wearable AR systems and information providers. Thereby, the information providers can provide users with the newest annotation information by updating the annotation database. On the other hand, users of AR systems can obtain the newest annotations without holding the annotation information in advance. The information providers can efficiently update and add the database with a web browser. Moreover, a wearable AR user can also edit the database of annotation information easily because the user's position acquired by a positioning sensor is used to determine the user's position on the map.

This paper is structured as follows. Section 2 describes the shared database system of annotation information using a wireless network. In Section 3, experimental results with a prototype system are described. Finally, Section 4 gives summary and future work.

2 Shared Database of Annotation Information

Figure 2 shows an outline of shared database system of annotation information. In this study, the database is shared via a wireless network. The database of annotation information is stored in the server and is shared by multiple users of wearable AR systems and information providers. Consequently, users of wearable AR systems can obtain annotation information at anytime via a wireless network and can see the newest annotation overlay images without holding the database of annotation information in advance. On the other hand, information providers can provide efficiently the newest annotation information for users of wearable AR systems by updating and adding the database with a web browser. In Section 2.1, the composition of the database of annotation information is described. Section 2.2 describes how to update the database with a web browser. Section 2.3 describes how the user obtains annotation information.

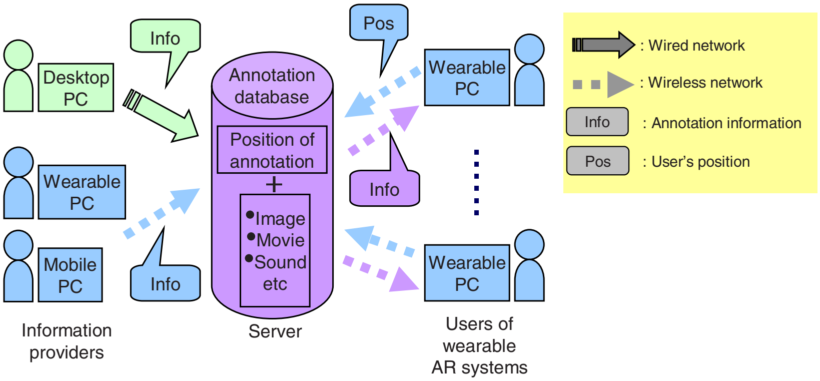


Fig. 2. Shared database of annotation information.

2.1 Composition of Annotation Database

The database contains some kinds of location-based contents. Each annotation is composed of a pair of contents(name and detail) and their positions. Components of the annotation information are described in detail in the following.

Position: Three-dimensional position of an annotation in the real world. Three parameters (latitude, longitude, height) are stored in the database.

Name: A name of the object which is overlaid in the real scene as annotation information.

Detail: Detailed information about the object. When user's eyes are fixed on the object, the detail about the object is shown in the lower part of the user's view.

2.2 Updating the Shared Database

The annotation information can be corrected, added and deleted by information providers with a web browser. An interface for information providers to update the database is a web browser as shown in Figure 3. Information providers can easily update the database by accessing a prepared web page and by transmitting the data of annotation information. The annotation updating procedure is described below.

1. **Specification of position**
   Information providers can zoom in and out to maps (Figure 3: C, D) using buttons (Figure 3: A). Besides, the providers can move by clicking any point on the map. In this way, the providers can specify the position of a new annotation to be added. It should be noted that position parameters such as latitude, longitude and height are automatically determined based on the specified position on the map.

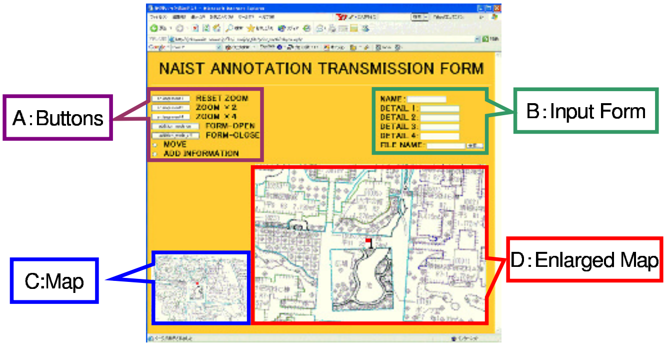


Fig. 3. Input form of annotation information.

2. **Input of name**
   Information providers input the object name to web page. The name is sent to the server and the picture of annotation is automatically generated in the server.
3. **Input of details**
   Using the same method as in the input of the name, information providers send details of objects. The providers also can send a picture, a sound file and a movie file as details.

The providers can efficiently send the newest annotation information using a web browser. For that reason, a user of wearable AR systems can also update the database. In this case, the server shows the user a map of his neighborhood according to the user's position acquired by positioning sensors. Since the user is able to update the database, the user can immediately correct the position error of annotations by confirming the overlaid image.

2.3 Getting Annotation Information

In this work, a database server is prepared assuming that the user's wearable computer can access the database via a wireless network. Annotations to be presented to the user are determined based on the user's current position. First, the user's position is measured by some sensors (positioning infrastructures, GPS, and so on) which are equipped by the user. The user's wearable computer then obtains proper annotation information based on the measured user's position. The server automatically decides which annotation should be provided. Consequently, the user's wearable system can obtain the newest annotation information at anytime. The user's wearable system obtains the newest annotation information periodically when the user moves for a fixed distance or a fixed time is passed.

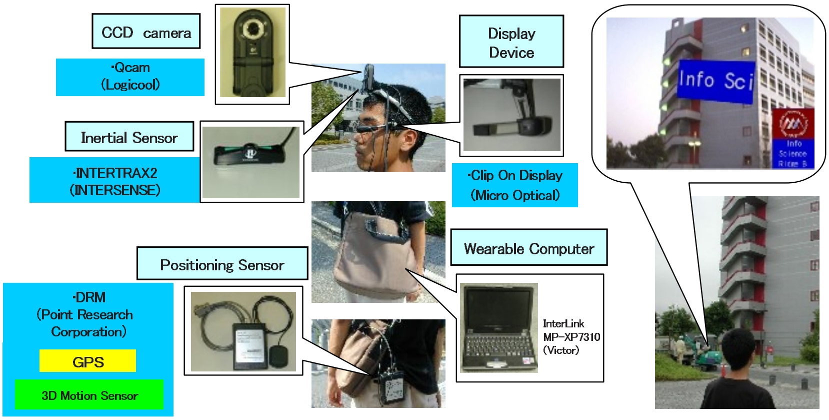


Fig. 4. Hardware configuration of wearable augmented reality system.

3 Experiments

We have carried out some experiments using the proposed database of annotation information in a server in our campus where users of wearable AR systems can use a wireless local area network. Figure 4 illustrates a hardware configuration of a wearable augmented reality system which is used in these experiments. The user equips some positioning sensors, a notebook PC and a display device. Three sensors described later can obtain the position and orientation of the user's viewpoint and the real scene image [12]. These data are sent to the notebook PC. The notebook PC obtains annotation information from the database server via a wireless local area network. The notebook PC sends annotation overlay images to a display device attached to the user's headset. The user can see it through the display device. Components of the system are described in more detail below.

Sensors. The user equips the following three sensors. Electric power is supplied from the notebook PC or a 9V battery. The data is transmitted to the computer through USB or serial connection.

Inertial sensor. (Intersense: InterTrax^2) The inertial sensor is attached to the user's headset and measures the orientation of the user's viewpoint. The inertial sensor can obtain data at 256Hz.

Camera. (Logicool: Qcam) The camera is attached to the user's headset and captures the real scene image from the user's viewpoint. It can capture a color image of 640 × 480 pixels at 30fps.

Positioning sensor. (Point Research Corporation: Dead Reckoning Module) The positioning sensor can measure the latitude and longitude. It can also measure accelerations in the horizontal direction.

Computer. (DELL: Inspiron8100, PentiumIII 1.2GHz, memory 512Mbytes) The computer is carried in the user's shoulder bag. It can use a wireless local area network with a network card.

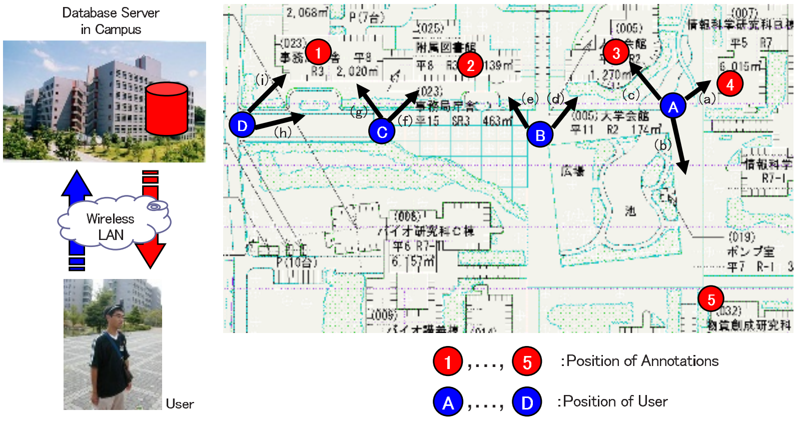


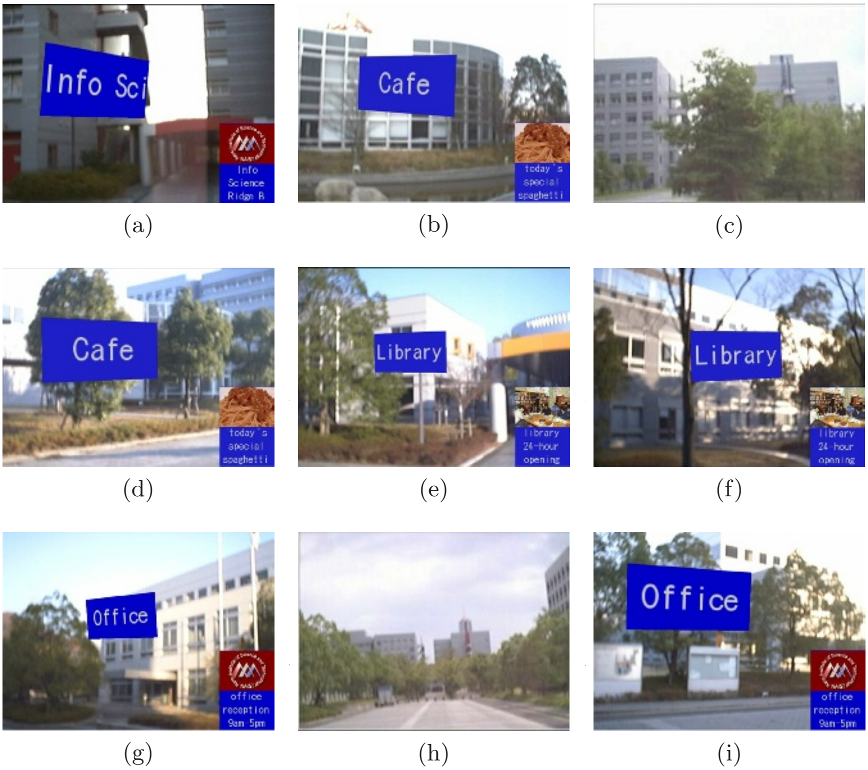
Fig. 5. Environment of the outdoor experiment.

Display device. (MicroOptical: Clip On Display) The display device is a video see-through device. It is attached to user's headset. It can present a 640 × 480 color image to the user.

In this experiment, we have developed a database of annotation information in the server (CPU Pentium4 2.0GHz, memory 512Mbytes) in our laboratory. Figure 5 illustrates the experimental environment. In Figure 5, the points 1,...,5 indicate the positions where the annotations exist. The user obtained the annotation information and looked around at the points A,...,D. The criterion concerning which annotations should be obtained is based on the distance between each annotation and the user's position. In this experiment, we set empirically the criterion at 70 meters. Besides, in order to check that the database is correctly updated, the user moving in our campus updated the annotation information about the "Cafeteria" with a web browser.

Figures 6 and 7 show the annotation overlay images. Figure 6 shows the annotation overlay images when the user was at the points A,...,D in Figure 5 and the user's orientation was along the arrows (a),...,(i) in Figure 5, respectively. As shown in Figure 6 (a), the annotation of "Information Science" is overlaid on the front of the building, so that the user can recognize the annotation information intuitively. The same conclusion is obtained from Figure 6 (b),...,(i). Thereby, we have confirmed that the user can obtained and perceive the shared annotation information intuitively. Figure 7 shows the annotation overlay images before and after updating the annotation information. The annotation information in Figure 7 (a) was changed to the new one in Figure 7 (b) automatically when a fixed time is passed. We have confirmed that the annotation information can be updated by editing the shared database of annotation information.

Through the experiments, the user has successfully obtained the location-based annotation information according to the user's position. Simultaneously, the shared database can be easily and efficiently updated and can provide the user with the newest annotation information in real-time.

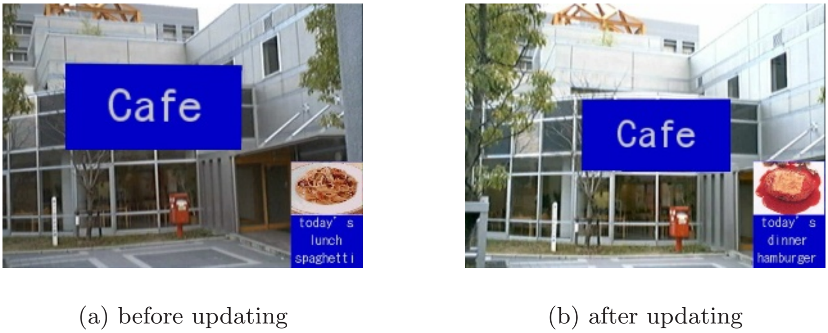


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 6. Overlay images at the points A, B, C, and D((a),(b) and (c) at A; (d) and (e) at B; (f) and (g) at C; (h) and (i) at D).



(a) before updating (b) after updating

Fig. 7. Example of updating annotation information.

4 Summary

This paper has described a database of annotation information for a wearable augmented reality system which is shared by multiple users via network and is efficiently updated with a web browser. In other words, proposed is a networked wearable augmented reality system. We have shown the feasibility of the proposed database through the demonstration with experiments in our campus. In the future, we should conduct experiments in a wider area and use other kinds of detailed location-based contents (movie, sound, and so on).

Acknowledgments. This work was supported in part by Core Research for Evolutional Science and Technology (CREST) Program "Advanced Media Technology for Everyday Living" of Japan Science and Technology Agency (JST).

References

1. S. Mann: "Wearable Computing: A First Step Toward Personal Imaging," IEEE Computer, Vol. 30, No. 2, 2002
2. R. Azuma: "A Survey of Augmented Reality," Presence, Vol. 6, No. 4, pp. 355–385, 1997
3. M. Kanbara, T. Okuma, H. Takemura and N.Yokoya: "A Stereoscopic Video See-through Augmented Reality System Based on Real-time Vision-based Registration," Proc. IEEE Int. Conf. on Virtual Reality 2000, pp. 255–262, 2000
4. S. Julier, M. Lanzagorta, Y. Baillot, L. Rosenblum, S. Feiner, T. Holler, and S. Sestito: "Information Filtering for Mobile Augmented Reality," Proc. 1st IEEE/ACM Int. Symp. on Augmented Reality, pp. 3–11, 2000
5. K. Satoh, K. Hara, M. Anabuki, H.Yamamoto, and H.Tamura: "TOWNWEAR: An Outdoor Wearable MR System with High-precision Registration," Proc. 2nd Int. Symp. on Mixed Reality, pp. 210–211, 2001
6. R. Tenmoku, M. Kanbara, and N. Yokoya: "A Wearable Augmented Reality System Using an IrDA Device and a Passometer," Proc. SPIE, Vol. 5006, pp. 478–486, 2003
7. J. Loomis, R. Golledge, R. Klatzky, J.Speigle, and J. Tietz: "Personal Guidance System for the VisuallyImpaired," Proc. Int. Conf. on Assistive Technologies, pp. 85–90, 1994
8. M. Kourogi, T. Kurata, and K. Sakaue: "A Panorama-based Method of Personal Positioning and Orientation and Its Real-time Applications for Wearable Computers," Proc. 5th IEEE Int. Symp. on Wearable Computers, pp. 107–114, 2001
9. M. Billinghurst, S. Weghorst, and T. Furness III: "Wearable Computers for Three Dimensional CSCW," Proc. 1st IEEE Int. Symp. on Wearable Computers, pp. 39–46, 1997
10. T. Okuma, T. Kurata, and K. Sakaue: "Fiducial-less 3-D Object Tracking in AR Systems Based on the Integration of Top-down and Bottom-up Approaches and Automatic Database Addition," Proc. 2nd IEEE/ACM Int. Symp. on Mixed and Augmented Reality, pp. 342–343, 2003
11. D. Stricker, J. Karigiannis, I. T. Christou, T. Gleue, and N. Ioannidis: "Augmented Reality for Visitors of Cultual Heritage Sites," Proc. Int. Conf. on Cultual and Scientific Aspects of Experimental Media Spaces, pp. 89–93, 2001

12. R. Tenmoku, M. Kanbara, and N. Yokoya: “A wearable Augmented Reality System Using Positioning Infrastructures and a Pedometer”, Proc. 7th IEEE Int. Symp. on Wearable Computers, pp. 110–117, 2003
13. A. State, G. Horita, D. Chen, W. Garrett, and M. Livingston: “Superior Augmented Reality Registration by Integrating Landmark Tracking and Magnetic Tracking,” Proc. SIGGRAPH’96, pp. 429–438, 1996
14. H. Petrie, V. Johnson, T. Strothotte, A. Raab, S. Fritz, and R. Michel: “MoBIC: Designing a Travel Aid for Blind and Elderly People,” Jour. of Navigation, Vol. 49, No. 1, pp. 44–52, 1996
15. S. Feiner, B. MacIntyre, T. Holler, and A. Webster: “A Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment,” Proc. 1st IEEE Int. Symp. on Wearable Computers, pp. 74–81, 1997

A Study on Image Electronic Money Based on Watermarking Technique

Jung-Soo Lee[1,2], Jong-Weon Kim[3], Kyu-Tae Kim[1],
Jong-Uk Choi[1,3], and Whoi-Yul Kim[2]

[1] MarkAny Inc., 10F, Ssanglim Bldg., 151-11, Ssanglim-dong,
Jung-gu, Seoul, 100-400, Korea
{jslee,jedam,juchoi}@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
[3] Col. of Computer Software and Media Tech., Sangmyung University,
7, Hongji-dong, Jongno-gu, Seoul, 110-743, Korea
{jwkim,juchoi}@smu.ac.kr

Abstract. This study introduces a technology that utilizes digital images as electronic money by inserting watermark into the images. Watermark technology assigns contents-ID to images and inserts the ID into the images in an unnoticeable way. The server that manages the issue and the usage of image electronic money (called 'WaterCash' hereafter) stores contents-IDs to database and manage them as electronic money. WaterCash guarantees anonymity and prevents the forgery and modification of WaterCash based on semi-fragile watermarking technique. In addition, WaterCash is transferable and the illegal use of WaterCash can be prevented based on the watermarking technology. Because the watermarking technology used in this paper was designed to be robust to image compression but vulnerable to intentional or unintentional image processing, WaterCash is applied to JPEG-compressed images.

1 Introduction

With the recent increase of cyber business such as Internet transactions and information services and the rapid rise of e-commerce, there are increasing demands for payment means that protect individual privacy and prevent forgery and modification [1,2]. Electronic money is electronic payment means that can be used without direct access to bank accounts, so they are being settled as efficient payment tools in e-commerce.

Electronic money, however, are always exposed to risks of duplication, forgery, modification, theft, etc. And the costs of issue and discard are high [2–4].

To solve these problems, we introduce new electronic money utilizing digital images as a bill. As embedding an image ID (content ID) to the image using the watermarking technology, forgery and modification of WaterCash is prevented. In addition, WaterCash provides transferability and anonymity concerning the

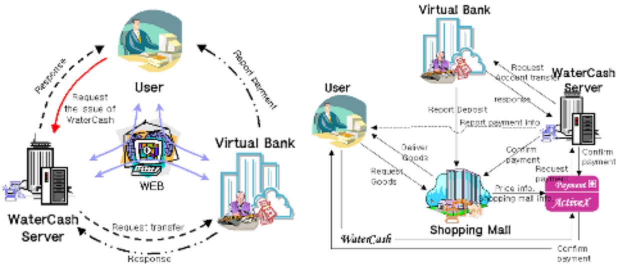
K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 508–514, 2004.

use of the bill for the payer's privacy [5–8]. What is more, as easily producible digital images are used as electronic money, WaterCash can save the huge amount of money for issuing and discarding. The semi-fragile watermarking technology used in this study is designed to cut off fundamentally the forgery and modification of WaterCash, and if it is forged or modified, its function is nullified [3,4].

The outline of this paper is as follows. Section 2 describes the general system of WaterCash including the structure of the system and the functions and roles of each component. Section 3 explains watermarking technology used in this study and the structure of data to be inserted into images. Section 4 presents how efficiently forged or modified parts on WaterCash are detected and measures how robust WaterCash is to JPEG compression. Finally, we give conclusions on the proposed e-money (WaterCash) and discuss its application in Section 5.

2 WaterCash System Structure

The Fig. 1 below is the general flow of the issue and the usage of WaterCash. It shows how WaterCash is issued and used as a payment means at shopping malls. The system is based on data that are inserted into images using watermarking technology.



(a) The issue of WaterCash (b) The usage of WaterCash

Fig. 1. The issue and usage of WaterCash.

2.1 The Issue and the Usage of WaterCash

Users can get WaterCash through WaterCash server. For getting WaterCash, he/she must input the account number of the payment bank. After WaterCash server checks if there is the balance as much as he/she requests in the account, if enough, issues WaterCash. Of course, money as much as he/she requests is transferred from his/her account to WaterCash server's account. And it records information about the issued WaterCash into database including the contents-ID and the date of issue.

To use WaterCash issued, the user saves it into a portable storage device or hard disk. A user who connects to an online shopping mall to purchase goods selects his/her WaterCash and inputs its password for payment. In the process, the user does not have to inform personal information other than data to receive the purchased goods. When the user inputs the password, the unique contents-ID is extracted from WaterCash and is sent to the WaterCash server. On receiving the contents-ID, the WaterCash server examines the validity of the WaterCash and, for valid WaterCash money, informs that payment has been made. In addition, the WaterCash server updates the database concerning the processed WaterCash and transfers the amount of payment to the account of the corresponding shopping mall.

2.2 WaterCash Transfer

In addition to functions described above, WaterCash can be handed over others. WaterCash is handed over only via the WaterCash server. That is, if the WaterCash server receives the contents-ID of WaterCash to be handed over, it disuses the contents-ID in database and inserts a new contents-ID into a new digital image and sends it to the transferee by email. The transferee who receives WaterCash changes the initialized password before using it.

3 Watermarking Technology and Data Structure

3.1 Watermarking Technology

In this study watermarking technology is used in inserting contents-ID and other information into WaterCash. This study utilized semi-fragile watermarking technique, which is designed to be robust to image compression but vulnerable to other malicious image modification. That is, if a part or the whole of WaterCash, in which data were inserted, is modified the watermark on the modified part is broken and consequently the WaterCash becomes invalid.

3.1.1 Dither Modulation
This section is devoted to explain the dither modulation. We divide section $\varDelta$ into n segments according to the quantity of data that we want to insert in each block and apply the dither modulation using Eq. (1).

$$D_c^{d_i}(u,v) = sign\left(C(u,v)\right) \times \left\{Q\left(|C(u,v)| + \frac{\varDelta}{2} - \frac{\varDelta}{n} \times d_i\right) + \frac{\varDelta}{n} \times d_i\right\} \quad (1)$$
$$u,v = 0,1,\cdots,7 \quad and \quad d_i = 0,1,\cdots,n-1$$

Here, $\varDelta$ is the length of a section of dither modulation. And $D_C^{d_i}$ is DCT coefficient after dither modulation. And $|C(u,v)|$ indicates the absolute value of the DCT coefficient in the position(u,v) to the dither modulation($u,v = 0,1,\cdots,7$), d_i is data to be inserted and $n = 2^k$ $(n \geq 2)$ and k is the number of bits inserted to each block.
sign (•) means the sign of input value, which is expressed as follows.

$$sign\,(x) = \begin{cases} x/|x| & if \quad x \neq 0 \\ 1 & if \quad x = 0 \end{cases} \quad (2)$$

In addition, $Q(\bullet)$ means the quantization of input value with $\varDelta$.

$$Q(x) = \lfloor x/\varDelta \rfloor \times \varDelta \tag{3}$$

Here, $\lfloor \bullet \rfloor$ rounds the result of operations to the nearest integers towards minus infinity.

For example, if $d_i = 1$ and $\varDelta = 32$ when $n = 2$ and DCT coefficient is 24, D_C^1 becomes '16'. If n is bigger, which means section $\varDelta$ is divided into more segments, then a large amount of data can be inserted. In this case, however, the robustness to compression is lowered. Accordingly, the quantity of data to be inserted and robustness to compression are in the relation of trade-off with each other.

3.1.2 Inserting watermark

This section introduces image watermarking technology used in inserting data into WaterCash. Signals used as watermark are contents-ID assigned to WaterCash, rewriting-prevention code and user password.

First, convert input data(contents-ID, rewriting-prevention code and user password) into binary codes of '0' and '1' and insert 1 bit into each 8x8 pixel block of the input image. The procedure of data insertion is as follows.
a. Perform DCT on an input image by 8x8 pixel block.
b. Perform dither modulation of the DCT coefficient using eq. 1 according to data (0 or 1) to be inserted.
c. Perform inverse DCT for DCT coefficient after dither modulation.

3.1.3 Extracting watermark

The process of watermark extraction is similar to that of watermark insertion. This section explains the process of extracting data that have been inserted as watermark.
a. Divide an input image into 8x8 pixel block and perform DCT.
b. Extract inserted data using eq. 4.

$$E_{d_i} = MOD_n\left\{\left[\frac{|C(u,v)|}{\varDelta/n}\right]\right\} \qquad E_{d_i} = 0, 1, \cdots, n-1 \tag{4}$$

Here, E_{d_i} means data extracted from the watermarked image. $MOD_n\{\bullet\}$ means the remainder after dividing the input value by n and $[\bullet]$ produces the closest integer to the input value.
c. Make the extracted data into meaningful codes.

3.2 The Structure of Inserted Data

To use images as electronic money we insert contents-ID into the images. What is more, we insert a rewriting-prevention code to prevent the contents-ID and user password from rewriting to WaterCash.

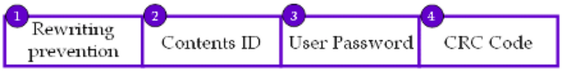


Fig. 2. The structure of data to be inserted.

When contents-ID is inserted to an image, it is checked whether contents-ID has already been inserted into the image before. If contents-ID has been inserted into the image, it is not allowed to replace with another contents-ID. Contents-ID area on the second field is to insert contents-ID assigned to an image. This contents-ID makes it possible to utilize WaterCash as money. That is, through verification procedure about the validity of the contents-ID in the WaterCash server, WaterCash can be used in every commercial transaction.

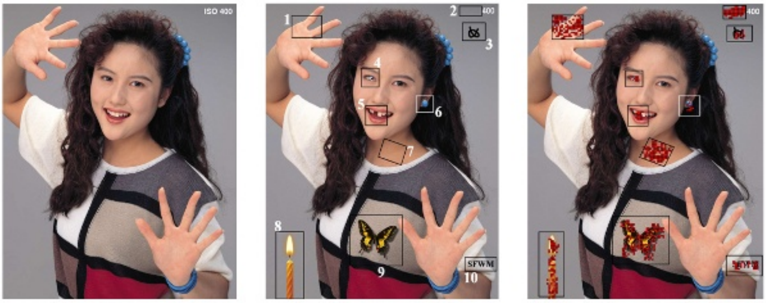
User password on the third field protects the user from damage by theft or loss. Because user password is decided through hash function, its length is not limited. Thus, once a password is set, other people are not allowed to use the WaterCash illegally. Lastly, cyclic redundancy check(CRC) code is attached at the end of these three fields. This CRC code makes it possible to verify whether the extracted data are correct or not.

4 Experiment Results

This section presents an experiment on the watermarking technique developed. Because when WaterCash is forged and modified it must be changed uselessly and images are generally saved in compression format we apply semi-fragile watermarking technique in this study. That is, to use the images as electronic money, proposed watermarking algorithm must be robust to compression but weak to other malicious image manipulation. If a user has modified a part or the whole of an image in order to change information on his/her WaterCash, the WaterCash becomes invalid. It is because the password becomes incorrect or contents-ID is not extractable due to the modified parts. This experiment tests whether the modified parts are efficiently detected when random data are inserted into images and parts of the images are modified. In addition, it includes that the extraction rate of the embedded information under JPEG-compressed image is shown.

4.1 Detecting Forgery / Modification

Let us assume that a malicious user has modified a part of an image used as WaterCash. And assume that we know data to be embedded for experiment. Through this experiment we verify whether the proposed watermarking algorithm can detect the modified parts. The size of the image used was 536 x 240, and a bit was inserted into each 8x8 pixel block. Figure 3 (a) shows the image to which data have inserted, and (b) is a partly modified image. (c) shows the detection of modified parts using a forgery/modification detection system. The modified parts can be detected through comparison the extracted data with



(a) Image Data inserted (b) Partly Modified Image (c) Detecting Modified Parts

Fig. 3. Forgery / Modification Detection. 1 : Paste original image, 2 : Delete (fill background textures), 3 : Add a line drawing, 4 : Change Color (the pupil of the eye), 5 : Delete guts (a tooth), 6 : Copy (an earring), 7 : Rotate, 8 : Paste another content, 9 : Paste another contents, 10 : Replace with computer generated texts.

the embedded data. But if modifications are actually applied to WaterCash, we decide whether the extracted data is correct or not through the CRC code.

4.2 Robustness to Compression

The table below is the results of experimenting on the robustness of the proposed semi-fragile watermarking technique to JPEG compression. It shows data extraction rates for different compression rates. In case of using 'BMP' or 'RAW' formatted image file as WaterCash, because the file size to be transmitted from WaterCash server to user is so large, JPEG-compressed images have to be used as WaterCash. So the proposed data embedding technique need to have robustness against image compression.

In the Table 1, QF means quality factor of the JPEG-compressed image. And extraction rate R_E is calculated using Eq. (5).

$$R_E = 1 - BER, \qquad BER = \frac{B_{Err}}{B_{Total}} \tag{5}$$

Here, B_{Total} is the total number of bits inserted, and B_{Err} is the number of wrong ones among extracted bits. And BER means bit error rate.

According to the graph, extraction rate goes down with the rise of compression rate. Because the value of WaterCash can be estimated properly when inserted data are extracted exactly, there should not be errors in data extracted from WaterCash. This problem will be solved using ECC(error correction code) because B_{Err} is under 10 %.

Table 1. Test on robustness to JPEG compression.

JPEG(QF) R_E	100	90	80	70	60	50	40	30
1-BER	1	0.997	0.995	0.994	0.986	0.987	0.985	0.968

5 Conclusions

In this study, we proposed the image electronic-money using the image watermarking technology. Because the technology is based on semi-fragile watermarking technique, it is possible to extract data from compressed images but not possible from maliciously modified images. In addition, because WaterCash prevents theft, forgery and modification and guarantees anonymity and transferability, it is safe and free from the invasion of personal privacy.

As it is getting easier to obtain and produce digital images, WaterCash using images is less expensive in issuing and discarding than currency or prior electronic money such as credit cards. What is more, as any kinds of images are usable as money, image electronic money may be utilized in advertising or as merchandise coupons.

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

References

1. Darius Buntinas, Eric Mazuk.: Digital Cash and Electronic Commerce. (1997).
2. Group of Ten.: Electronic money. http://www.bis.org/publ/gten01.htm, April (1997), ISBN 92-9131-901-5.
3. M. U. Celik, G. Sharma, E. Saber, A. M. Tekalp.: Hierarchical Watermarking for Secure Image Authentication wit2h Localization," IEEE Trans. Image Proc., Vol. 11, No. 6, June (2002).
4. J. J. Eggers, R. Bauml, B. Girod.: A communications approach to image steganography. In SPIE Electronic Imaging 2002, Security and Watermarking of Multimedia Contents IV, vol. 4675, San Jose, USA, Jan. (2002) 26-37.
5. J. Carnenisch, J-M. Piveteau, and Stadler M.: An efficient electronic payment system protecting privacy. In Computer Security ESORICS'94(LNCS 875), Springer Verlag, (1994) 207-215.
6. G.Davida,Y.Frankel,Y.Tsiounis,M.Yung.: Anonymity Control in E-Cash Systems. Financial Cryptography'97, (1997).
7. G. Davida, Y. Frankel, Y.T. siounis, M. Yung.: Anonymity Control in E-Cash System. Proc. Financial Cryptography Workshop, Feb., (1997).
8. F. Zhang and K. Kim.: ID-based blind signature and ring signature from pairings. Advances in Cryptology-Asiacrypt 2002, (2002) 354-368.
9. S. Kim and H, Oh.: A new electronic check system with reusable refunds. Inst. J. Information Security, vol.1, no.3, (2002) 175-188.
10. M. Puhrerfellner.: An implementation of the Millicent micro-payment protocol and its application in a pay-per-view business model. Master's thesis, Distributed Systems Group, Technical University of Vienna, Austria, (2000).

A Fully Automated Web-Based TV-News System

P.S. Lai, L.Y. Lai, T.C. Tseng, Y.H. Chen, and Hsin-Chia Fu[*]

Department of Computer Science and Information Engineering,
National Chiao-Tung University
Hsinchu 300, Taiwan
{pslai,lylai,cltseng,yuehhong,hcfu}@csie.nctu.edu.tw

Abstract. This paper proposes a web-based multimedia TV-News system, which can records and analyzes news video to generate hierarchy news contents automatically. To achieve this goal, various neural network based techniques, including image processing, audio processing, and optical characters recognition are applied. Since July 2003, the multimedia TV-News system has been implemented and continuously been up running at http://nn.csie.nctu.edu.tw/TVNews/intro.htm for general public browsing and studying.

1 Introduction

With the emerging of network service, web-news[1][2] becomes a popular information resource for people to make contact with the world. The most significant characteristics of web-news service are instant data updating, hierarchy news contents, and user-friendly searching mechanism. However, lacking of video is a drawback comparing with traditional TV-news programs. Although TV-news program has colorful videos, it is still painful for users to find stories they really want. Due to the recent advances of web technology on multimedia, creating a news media that possesses all the advantages of web-news and TV-news becomes possible. Such a web-based-service can provide a well-organized daily news list, convenient searching mechanism, and rich multimedia contents. And most importantly, a system that can generating contents fully automated. Browsing headline list may be the simplest way for users to access news stories. Therefore, making title for each news story is the first thing to do. Then, in order to efficiently achieve a clear idea of a news story without watching the whole video clip, a set of key frame images can be used to depict the story. In general, headlines, key-frames, video clips, and story scripts are needed contents for a multimedia news service.

Using digital multimedia techniques to create TV news programs has been a new trend for news media production system. However, TV news has been broadcasted for years, a lot of TV news contents are saved and preserved in

[*] This research was supported in part by the National Science Council under Grant NSC 91-2213-E009-115.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 515–522, 2004.

thousands of news videotapes. Thus, an automatic hierarchy news generating system is definitely necessary to produce multimedia contents from these tapes. Although there are difficulties to retrieve contents from video, the maturation of multimedia and pattern recognition techniques signals we are now able to conquer all the problems.

In general, shot detection[3], speaker identification[4], video optical character recognition(VOCR)[5][6], and data mining[7][8] techniques are needed for video analysis and multimedia content generation. Several multimedia systems[9] have been proposed. Most of which are designed for alphabetical language based media. Here, we propose to use the VOCR techniques to recognize Chinese caption for extracting key words in a news story. The difficulties in VOCR for poor character image are due to 1) many characters(5401 words), and 2) complicated patterns. To tackle these problems, we propose a frequency-based VOCR technique and searching related text over web news to improve the recognition accuracy. In order to demonstration the proposed system, a prototype of the multimedia web system is available at http://nn.csie.nctu.edu.tw/TVNews/intro.htm. This paper presents a web-based TV-news system that generates contents automatically. The general system architecture is framed in Section 2. The details of how necessary data are generated automatically is illustrated in Section 3. And, Section 4 discusses the design and implementation of the web-based user interface. Finally, Section 5 briefs concluding remarks and future works.

2 System Architecture

The flow chart of automatic news content generation is depicted in Figure 1. There are two input sources - Cable TV and World Wide Web. At first, TV-news program video is recorded to produce high-quality video for analysis, and to generate streaming video for web browsing. The news-video is fed into modules for story segmentation, key-frame selection and headline generating. Then, news story headlines, key frames, streaming video clips, and scripts are stored in a database. We will discuss these technologies in detail in the following sections.

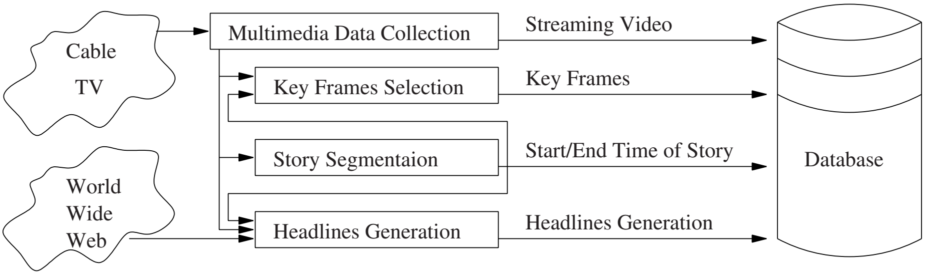


Fig. 1. Flow chart of automatic news content generation. There are two input sources, including Cable TV and World Wide Web, and four output contents, including streaming video, key-frames, start/end time of story, and headlines.

A user could requests through the proposed TV-news service web site to search story by keywords or browse daily news. The detail of web-interface design will be given in Section 4.

3 Content Analysis

Headlines, and key-frames of TV news stories, and anchors' speech models can be obtained from video content analysis. The content analysis consists of four components: multimedia data acquisition, key-frame extraction, story segmentation, and news headline generation. The details are recorded as follows.

3.1 Multimedia Data Acquisition

In order to automatically generate necessary contents, news video and scripts are collected at the beginning of the work. Then, the captured data is transformed into suitable format before being applied to the following analysis. For general content analysis work, MPEG-1, a well-defined open standard of fair quality video, is the format we need the most. Except MPEG-1 video, high-quality and well-recorded images are needed for close-captions extraction and VOCR. For this purpose, the captured TV-frames are sampled into portable pixel map (PPM) images. Besides, the system encodes captured news video into ASF format for transiting video over Internet of various bandwidths. The advancement of computer hardware makes capturing and encoding video into three different video-formats at the same time to be possible. In addition, Encoding video into all desired formats simultaneously brings many of advantages, such as processing time efficient and event synchronizing accurate, etc.

A robot like, web searching software was also developed to automatically fetches news scripts from net-news web sites.

3.2 Key-Frame Extraction

To efficiently browse news story without downloading a whole news video, a set of key frames are selected from sampled video images. The main idea is that the system firstly cuts video into several series frame sets, named shots, and then picks frames from each shot. The spatio-temporal slice method, proposed by Ngo[3] presents the spatio and temporal relationship of video sequences. Because the shot-change brings clear edges in spatio-temporal slice, the shot change locations can be easily detected by conventional edge detection algorithm. An example of spatio-temporal(ST) slice is shown in Fig 2. Two apparent vertical edges divide ST slice into three pieces. These two vertical lines corresponds to the time lines of shot changes.

To catch motion activities of a news story as much as possible, we extract key-frames from each shot, and high motion scenes.

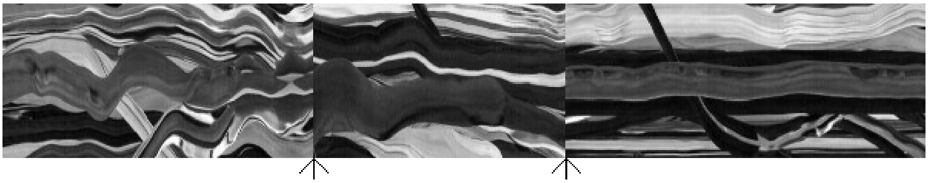


Fig. 2. An example of spatio-temporal slice. The locations indicated by arrow symbols are just the shot-change locations.

3.3 Story Segmentation

In the way of a traditional TV news presentation, news abstract is firstly narrated by an anchor man, and then the display of background story. Presentation of every news story is repeatedly in this manner. In the proposed method, this scenario is adopted to segment a daily news program into individual stories. To locate the proper time point when an anchor starts the narrative speak of a video clip, we use *Bayesian Information Criterion* based method [10] to locate the changing points of the speech. The changing point indicates the start and the end of individual segment. Segments with similar characteristic are linked together as a cluster. In general, anchors' speech segments forms the biggest cluster. Modeling the speech of the anchor with Gaussian mixture model and fine tuning by EM algorithm can further improve the preciseness of changing points between speakers. By scanning through the whole audio recording over the taped video, the starting time points of an anchor segment can be located accurately. The details of this method can be found in [4,11].

3.4 Headlines Generation

The section presents the title and descriptions extraction of a news story. TV News introduce each story in a few words by displaying them as close-captions. Extracting and recognizing close-captions from video could definitely achieve the title words. Unfortunately, low-resolution video character patterns causes the poor performance of optical Chinese character recognition. Therefore, to find supporting source for headline generating is necessary. Figure 3 shows an example to depict why video optical Chinese character recognition is difficult. Since the Chinese character is much complicated than alphabetical characters, thus Chinese character are not displayed clearly in low-resolution image, comparing to a printed Chinese caracter image. In addition, after binarization process, the character image becomes even worse.

The good news is that most TV stations provide text news headlines and scripts on their web site before the news programs are on the air. If we can find out the corresponding relation between news scripts of web-news and segmented video clips, the perfect news headlines can be achieved. Therefore, the system run video-OCR firstly, then matches extracted characters with web-news scripts

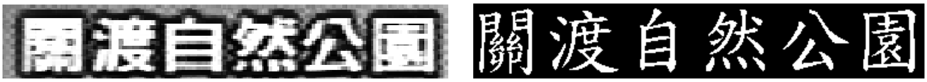


Fig. 3. An example of Chinee characters images. When comparing with the printed images(right), The quality of TV-news program frames(left) are very poor.

to link video clips and scripts. In addition to headlines, some scripts of a news story are linked, too.

In order to extract characters from frames, the system, firstly, detects the region of close-captions, and then segments individual character from the caption block.

A *two-layered* OCR-engine is designed to recognize these extracted character blocks. The images blocks are classified into several *coarse* classes, at first. Then, the characters are recognized by the *fine* classifier. As shown in Table 1, the low frequency part of discrete cosine transform performs best as the feature of coarse classifier. Two dimensional discrete cosine transform of an image of size N_1xN_2 is defines as

$$\hat{I}(x,y) = cu \cdot cv \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} I(i,j) cos\Big(\frac{\pi x(2i+1)}{2N_1}\Big) cos\Big(\frac{\pi y(2j+1)}{2N_2}\Big), \quad (1)$$

where $I(x,y)$ is the image intensity at location (x,y).

Table 1 also suggests that the best feature for fine classifier is generated by Daubechies wavelet transform. The scaling function (Eq. 2) and wavelet function (Eq. 3) of Daubechies D4 wavelet transform is defined as

$$a_i = \frac{1+\sqrt{3}}{4\sqrt{2}} s_{2i} + \frac{3+\sqrt{3}}{4\sqrt{2}} s_{2i+1} + \frac{3-\sqrt{3}}{4\sqrt{2}} s_{2i+2} + \frac{1-\sqrt{3}}{4\sqrt{2}} s_{2i+3} \quad (2)$$

$$c_i = \frac{1-\sqrt{3}}{4\sqrt{2}} s_{2i} - \frac{3-\sqrt{3}}{4\sqrt{2}} s_{2i+1} + \frac{3+\sqrt{3}}{4\sqrt{2}} s_{2i+2} - \frac{1+\sqrt{3}}{4\sqrt{2}} s_{2i+3} \quad (3)$$

where s_i is the image intensity.

Table 1. Experimental results of feature combinations. 21711 training and 13384 testing samples of 1446 characters are used. Six features, including closing count(CC), first order peripheral(FPF), text pixel distribution(TPD), central projection(CPT), discrete cosine transform(DCT), and discrete wavelet transform(DWT), ,are tested, and the best combination is DCT+DWT(94.37%).

coarse\fine	CC	FPF	TPD	CPT (128)	DCT (210)	DWT (256)
CC(32)	90.17%	79.84%	$< 50\%$	89.79%	93.84%	93.73%
FPF(64)	90.47%	80.22%	93.16%	90.80%	94.18%	94.13%
TPD(64)	89.89%	77.47%	91.28%	87.86%	93.46%	93.27%
CPT(128)	90.77%	79.94%	92.63%	91.04%	93.33%	92.63%
DCT(55)	91.63%	80.32%	$< 50\%$	91.11%	94.24%	94.14%
DWT(64)	91.55%	81.15%	93.66%	91.30%	94.37%	94.19%

After all features are extracted, the k-means algorithm is applied to partition characters into clusters. Efficiency is the major reason. The k-means algorithm repeats the following steps until all clusters centers are converged.

1. Redistribute data points to the nearest cluster k_i.

$$k_i = \underset{k}{\operatorname{argmin}} \frac{1}{V} \sqrt{\sum_v (p_{iv} - c_{kv})^2} \tag{4}$$

   where V is the number of features, p_{iv} is the v-th feature of point p_i and c_{kv} is the v-th feature of cluster center c_k.
2. Compute the center of cluster $c_k = (c_{k1}, c_{k2}, .., c_{kV})$

$$c_{kv} = \frac{1}{N_k} \sum_{i \in C_k} p_{iv} \tag{5}$$

   where N_k is the number of data points belong to cluster C_k.

Finally, PAT-Tree[8] is used to match the extracted close-caption characters to scripts of web-news to obtained headline for each story.

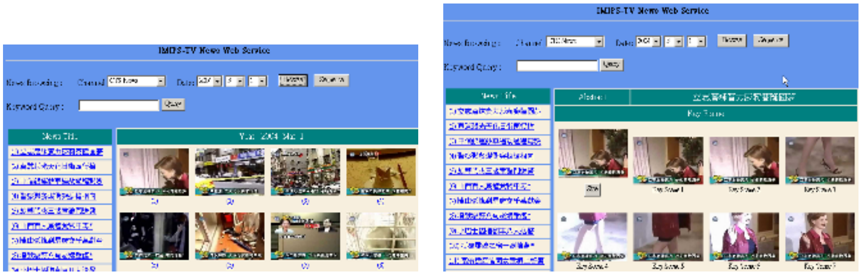
3.5 Discussion

The overall precision rate is infected by several factors. The first is that news channel often changes their presentation style. For example, all commercials were clustered together years ago, but recently commercials are spreaded over the whole news program. The second, fetching proper text documents from web is another dominating factor. In other words, the completeness of website's news story greatly affects the performance. To keep our system acceptable to meet news web user's requirement, we would like to modify the system to achieve 90% precision rate with 20% reject of each news story.

4 Web-Based User Interface

This section presents the design and implementation of web-base user interface. First, we would like to briefly describe three servers in the web TV-News system. These servers are database server, web server, and media server. The data generated in Section 3 are stored in the SQL database. Web server accepts users' requests of looking up contents from database server, and then composes and returns the requested pages to users. When users acquire a news video, web server redirects the request to media server. The media server then supplies required video.

At the prototype web site [12], users can browse news stories by date, or query desired news story by assigning keywords. After opening the starting page, there are two links for PC-users and PDA-users respectively to begin the news service. The following explanation is for PC users. The main service page is



(a) News stories list

(b) Key-frames and video of story

Fig. 4. User interface Demonstration.

divided into two partitions - top and bottom frames. Users can specify date and channel on the top frame, and browse news stories list of specified date and channel on the bottom one, as shown in Figure 4(a). Headlines are listed on the bottom-left. And the images displayed on the bottom-right are the representative key-frame of each news story. Users can select stories by clicking on headlines or representative key-frames. The story selected is shown on the bottom frame (see Figure 4(b)). The key-frames of this story are presented on the bottom-right all at once. Besides, a embedded window for playing video is also displayed on the top-left corner.

In addition to browsing related news stories by date and channel, users can assign several keywords for related news stories in the database of the proposed web TV-News system. Stories that matched users' requirements are then listed at the bottom-left frame.

5 Conclusion and Future Works

A web-based TV-news service system with automatic content generation function is proposed in this paper. The proposed system can be implemented on two PCs. The P4-1.8G-Hz machine is recommended to be used to generate contents, and the other one (P3-650MHz) serves as a database and a web server. Including the recording time, one-hour news video can generate contents in three hours. The prototype TV news browsing system was finished in July 2003. Since then, we powered up the system, and let in runs all the time. Up to the date of May 28, 2004, when the report is written, the system is continuously running, except a few short shutdown due to power failure.

The overall precision rate is infected by several factors. The first is that news channel oftenly changes their presentation style. For example, all commercials were clustered together years ago, but recently commercials are spreaded over the whole news program. The second, fetching proper text documents from web is another dominating factor. In other words, the completeness of website's news story greatly affects the performance. To keep our system acceptable to meet news web user's requirement, we would like to modify the system to achieve 90% precision rate with 20% reject of each news story.

However, the system could be improved in the following aspects: first, higher accuracy rate in headline generation and story segmentation are need to achieve better automated processing performance. second, more semantics meaningful key frames are needed. Knowledge discovering algorithm may be a better choice to generate more meaningful key-frames.

References

1. http://www.cna.com.tw
2. http://www.cnn.com
3. Ngo, C.W.: Analysis of Spatio-Temporal Slices for Video Content Representation. PhD thesis, The Hong Kong University of Science and Technology (2000)
4. Cheng, S., Chen, Y., Tseng, C., Fu, H.C., Pao, H.: A self-growing probabilistic decision-based neural network with applications to anchor/speaker identification. In: Proceedings of the Second International Conference on Hybrid Intelligent Systems (HIS'02), Santiago, Chile (2002)
5. Sato, T., Kanade, T., Hughes, E.K., Smith, M.A.: Video optical character recognition for digital news archive. In: Proc. Workshop on Content-Based Access of Image and Video Databases, Los Alamitos, CA (1998) 52–60
6. Fu, H.C., Chang, H.Y., Xu, Y.Y., Pao, H.T.: User adaptive handwriting recognition by self-growing probabilistic decision-based neural networks. IEEE Transactions on Neural Networks **11** (2000)
7. Chen, K.J., Liu, S.H.: Word identification for mandarin chinese sentences. In: Proceedings of the Fifteenth International Conference on Computational Linguistics, Nantes (1992) 101–107
8. Chien, L.F.: Pat-tree-based keyword extration for chinese information retrieval. In: the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA (1997) 50–58
9. http://www.informedia.cs.cmu.edu/
10. Cheng, S.S.: Model based learning for gaussian mixture model and its application on speaker identification. Master's thesis, The National China-Tung University (2002)
11. Y.H.Chen, C.L.Tseng, S.S.Cheng, T.M.Fang, H.Y.Chan, Fu, H.C.: On the scene classification for the automated generation of hierachical contents from broadcasting tv news. In: Proceedings of KES'02, Milan, Italy (2002)
12. http://nn.csie.nctu.edu.tw/tvnews/intro.htm

An Evolutionary Computing Approach for Mining of Bio-medical Images

Shashikala Tapaswi[1] and R.C. Joshi[2]

[1] Madhav Institute of Technology and Science, Gwalior, India
[2] Indian Institute of Technology, Roorkee, India

Abstract. The key requirement in medical imaging systems is to be able to display images relating to a particular disease, there is increasing interest in the use of Image Retrieval techniques to aid diagnosis by identifying the region of abnormalities from bio-medical images. Bio-medical images, such as pathology slides, usually have higher resolution than general-purpose pictures. In this paper an evolutionary computing based technique for classification of biomedical images on the basis of combined feature vector, which combines color and texture feature into a single feature vector, is presented. The system uses concept based on pixel descriptors, which combines the human perception of color and texture into a single vector, with the extraction of region of interest. The region extracted using the feature vectors represented in the form of pixel descriptor are fed as input to a neural network, which is trained for classification of images using genetic algorithm. The technique has been implemented on the database of biomedical images. Some of the experimental results are reported in the paper. The medical community can be assisted with this technique in diagnosing the disease.

1 Introduction

With the increase in modern medicine and diagnostics techniques such as radiology, histopathology, and computerized tomography has resulted in an explosion in the number and importance of medical images now stored by most hospitals. As more and more images are captured in electronic form the need for programs which can find region of interest in a database of images is increasing. While the prime requirement for medical imaging systems is to be able to display images relating to a named patient, there is increasing interest in the use of Image Retrieval techniques to aid diagnosis by identifying similar past cases. Probably the greatest medical advance in the late twentieth century was the development of CT scanning techniques, which in many instances removed the need for exploratory surgery [1,14]. The same CT techniques that make image reconstruction possible using X rays have subsequently been applied to magnetic resonance imaging, a more sensitive technique for analysis of soft tissue and for metabolic studies. The recent development of digital radiography is replacing traditional methods of storing X-ray film, with direct computer storage providing the ability to transfer images from the office to the physician's home or to remote locations.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 523–532, 2004.

In medical imaging it has been observed that the use of visual texture conveys useful diagnostic information. However image-processing modes based on scan sections or radiographic views do not completely provide diagnostic information in advance, when it would be easier to control a disease, make a therapeutic decision, or perform surgery. This is due to the fact that gray level differences in tissues are small compared to the accuracy with which the measurements may be carried out for a reasonable patient dose of X-rays. As there are many texture analysis methods available, it is possible to derive numerous texture parameters from a region of interest in an image. Image retrieval is critically important in patient digital libraries, clinical diagnosis, clinical trials and pathology slides. Most of the existing image retrieval systems [2,3,5,6] are designed for general-purpose picture libraries such as photos and graphs. Regardless of the imaging technology, all digitized images use the same general format. It is very important to extract the maximum possible information from any image obtained. Most of the image retrieval systems use low-level features such as color, texture, structure and shape. It is generally accepted that texture and color are the key features for image retrieval systems. There have been attempts to combine color and texture [2]. The paper uses the assumption of pixel descriptor [15], which encompasses color information into texture features. The pixel descriptor encodes color information that is with in human perception range [4,8,10]. The feature derived from pixel descriptor is more meaningful to human perception than the texture feature representation alone and significantly improves the retrieval performance. The performance of the proposed technique in which the features are represented as pixel descriptor are compared with Gabor texture representation [11,12]. The prominent regions are extracted using clustering techniques [19]. The obtained feature vector is used for training the neural network [7] for classification of biomedical images. This approach aims to use genetic algorithms to train and refine feature-based networks for the Region of Interest detection problems. The training of neural networks is done using genetic algorithms. The proposed technique can assist the medical community in diagnosing the disease. The paper has been organized as follows Section 2 explains the proposed evolutionary computing approach. The implementation of technique is discussed in Section 3, experimental results are presented in Section 4, and conclusion and future discussions are given in Section 5.

2 The Proposed Technique

Image Retrieval Systems, which exists today, are essentially limited by the way they function. Such as an Optical Character Recognition method may be good for graphs or charts found in biomedical educational area while a region-based approach is much better for pathology and radiology images. A method based on classification of images based on texture and color is presented for searching biomedical images. A combined feature vector for texture and color is obtained which is more meaningful than the texture feature representation alone and significantly improves the performance. The significant regions from images are

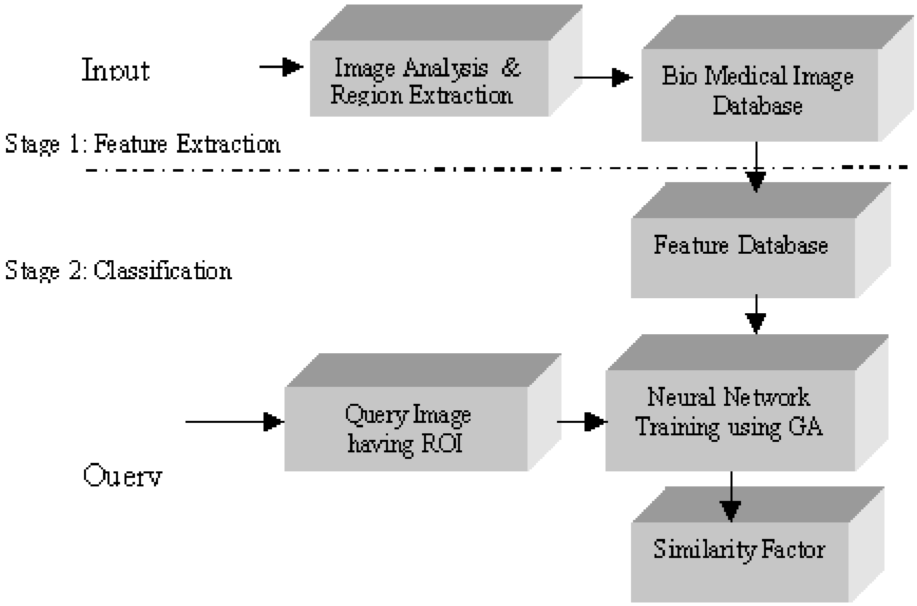

Input
Image Analysis & Region Extraction
Bio Medical Image Database
Stage 1: Feature Extraction
Feature Database
Stage 2: Classification
Query Image having ROI
Neural Network Training using GA
Query
Similarity Factor

Fig. 1. Block Diagram of the proposed technique

extracted and stored in the database. The method can be useful in diagnostic of diseases if the possibility of a color and texture is known in the Region of Interest (ROI) of an image, and the images can be classified into category of diseases. The neural network is trained using genetic algorithm. The output obtained after training the neural network is used to calculate the similarity factor [16]. Figure 1 gives the block diagram of the technique.

3 Implementation of the Technique

The technique implemented is carried out in two stages. First stage is the region extraction stage and second stage is of t he classification of images.

Stage 1: Region Extraction

For extraction of feature the concept of pixel descriptor [15] is used for describing the details of pixels in an image. The purpose of pixel descriptor is to combine the vector form of texture feature representation, derived from the responses to the image from a set of filters, with the color information. Pixel descriptors combine the texture information with color and are derived by using a set of texture and color descriptors. A Gabor function and a Gaussian function as a color descriptor function has been considered. A pixel descriptor for a pixel *(x,y)* in the image I is then obtained by taking a vector of both color and texture descriptors for that particular pixel (x, y).

3.1 Obtaining Pixel Descriptors

A Gabor function is used to derive a set of filter banks for texture description. Gabor functions are Gaussian functions modulated by complex sinusoids. In two dimensions, the Gabor functions are as follows:

$$g(x,y) = \frac{e^{\frac{-1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}) + 2\pi j\omega}}{2\pi\sigma_x\sigma_y}$$

where $j = \sqrt{-1}$, ω is the frequency of sinusoid and σs are standard deviations (parameters of the Gabor function). A class of self-similar Gabor wavelets by appropriate dilations and rotations of $g(x, y)$, through the generating function can be obtained.

$$g_{mn}(x,y) = a^{-m}g(x',y'),$$

$$a > 1, m, n = \text{int}\, eger,$$

$$x' = a^{-m}(x\cos\theta + y\sin\theta),$$

$$y' = a^{-m}(-x\sin\theta + y\cos\theta)$$

where $\theta = n\pi/K, K$ is the number of orientations and $n = 0, 1, ..., S-1, S$ is the number of scales. Let U_l and U_h denotes the lower and upper center frequencies of interest. Then the following filter design ensures that the half - peak magnitude support of the filter responses in the frequency spectrum touch each other.

$$\sigma_u = 1/2\pi\sigma_x, \quad \sigma_y = 1/2\pi\sigma_y,$$

$$a = \left(\frac{U_h}{U_l}\right)^{\frac{-1}{(s-1)}}, \quad \omega = U_h,$$

$$\sigma_u = \frac{((a-1)U_h)}{(a+1)\sqrt{2\ln 2}}$$

$$\sigma_v = \tan\left(\frac{\pi}{2K}\right)\left[U_h - 2\ln 2\left(\frac{\sigma_u^2}{U_h}\right)\right]\left[2\ln 2 - \frac{(2\ln 2)^2\sigma_u^2}{U_h^2}\right]^{\frac{1}{2}}$$

For experimental results reported, $S = 4$ and $K = 6$ and a filter size of 61×61 have been used. Given an Image $I(x, y)$, the transform coefficients are computed.

$$T_{mn} = \int\int I(x_1,y_1)g_{mn} * (x - x_1, y - y_1)dx_1dy_1$$

where $*$ indicates the complex conjugate. The texture descriptor for the given image I is then the vector of matrices,

$$u_{texture} = [T_1, T_2, ..., T_N] \quad \text{where} \quad N = S * K = 24.$$

A normalized Gaussian model to extract color descriptors of the images is used. The Gaussian model in n-dimension is represented as:

$$G_n(x) = K_c exp\left(-\frac{1}{2}\left(\frac{\sqrt{\sum_{i=1}^{n}(x_i-\mu_i)^2}}{\sigma}\right)^2\right)$$

In an image the total number of colors present is very large. There are over 16 millions colors in an image with 24 bit color. It is extremely difficult to model such a large number of colors. To limit the number of colors without loosing much information contained in an image from the point of user's perception, only 27 colors are chosen following the experimental results in [9,10,13]. It has been found experimentally that these 27 different colors are significant and fall within the range of human perception [9,13]. Thus in order to represent color descriptors, 27 different Gaussian functions are generated by changing the value of μ_i. In the RGB color space, the 3-dimensional ($n = 3$) Gaussian functions are formulated. μ represents the RGB values for a particular color. Suppose C_c is the response of the image I to a color c, the color descriptor for I is then given by the vector of matrices,

$$u_{color} = [C_1, C_2, ..., C_{27}]$$

By varying the value of σ the response of the Gaussian function can be tuned. For the experiments the values for $K_c = 1$, and $\sigma = 0.5$ are taken.

Pixel Descriptors are obtained from texture and color descriptors. The pixel descriptor represents the color descriptors and the texture descriptors. In the technique 24 texture descriptors and 27 color descriptors are used. For a pixel I at (x, y) in an image, the pixel descriptor is derived as follows :

$$u_{pixel}(x,y) = [u_{texture}(x,y), u_{color}(x,y)]$$

where $u_{texture}(x,y)$ is the 24 dimensional vector of $T(x,y)$ and similarly for color. Then the pixel descriptor vectors are normalized as follows.

$$u_{pixel}(x,y) = \frac{u_{pixel}(x,y)}{\|u_{pixel}(x,y)\|_2}$$

The subscript 2 stands for Euclidean distance. The above mentioned normalized pixel descriptor is used to generate feature representation.

A texture region is represented by mean μ and the standard deviation σ of the energy distributions of the transform coefficient, calculated as follows:

$$\mu = \int\int \mid u_{texture}(x,y) \mid dxdy$$

$$\sigma = \sqrt{\mid (u_{texture}(x,y) - \mu)^2 \mid d_x d_y}$$

Then the feature representation for texture descriptor alone is represented as follows.

$$r_{11} = [\mu_0, \sigma_0, ..., \mu_{23}, \sigma_{23}]^T$$

Similarly the feature representation from pixel descriptors are extracted as follows

$$\mu = \int\int \mid u_{color}(x,y) \mid dxdy$$

$$\sigma = \sqrt{\int\int \mid (u_{color}(x,y) - \mu)^2 \mid dxdy}$$

Then the feature representation for texture descriptor combined with color descriptors is represented as follows.

$$r_{12} = [\mu_0, \sigma_0, ..., \mu_{23}, \sigma_{23}, ..., \mu_{50}, \sigma_{50}]$$

For region extraction, these pixel descriptors are then clustered by using the clustering technique [19]. For clustering a constant number c of well-scattered points in a cluster are chosen first. The chosen scattered points are next shrunk towards the centroid of the cluster by a fraction α. The value of α is assumed as 0.3. It has been observed experimentally the value 0.3 for α generated right clusters and the effects of outliers were dampened. The scattered points after shrinking are used as representatives of the cluster. The clusters with the closest pair of representative points are grouped at each step. This clustering algorithm is less sensitive to outliers since shrinking the scattered points toward the mean dampens the adverse effects of outliers as these are typically far way from the mean. The number of clusters depends on the complexity of the image. It identifies a region of the image with related pixel values. The clustered pixels are sent back if necessary to the clustering module. The final module smoothes the regions and outputs the result.

Stage 2: Classification

In stage 2 the Region of Interest (ROI) is selected and is submitted as a query image to the neural network. The regions are extracted and stored in the medical image database using the feature based descriptor. The query image having abnormalities can be submitted to the Neural Network and training can be performed. Genetic algorithm is used for neural network training as with genetic algorithm, more accurate results are achieved. It would be beyond the scope of this paper to explain the basic terminology, implementation and features of genetic algorithm. In order to get an appropriate chromosomal representation of the network weights [17,18] these have to be randomly initialized multiple times and accordingly coded into linear structure. Each chromosome (individual) represents one neural set. Since the architecture of the neural network is predefined and remains fixed after the initialization the chromosome solely consists of the weight values does not contain any topological or structural information. All these individuals represent the initial population. For this kind of set up, the groups of weights are swapped over. Two population examples are selected, one from the lower error (LE) half, and another from the higher error (HE) half. The LE weights are given HE for the final layer, and put it back in the population.

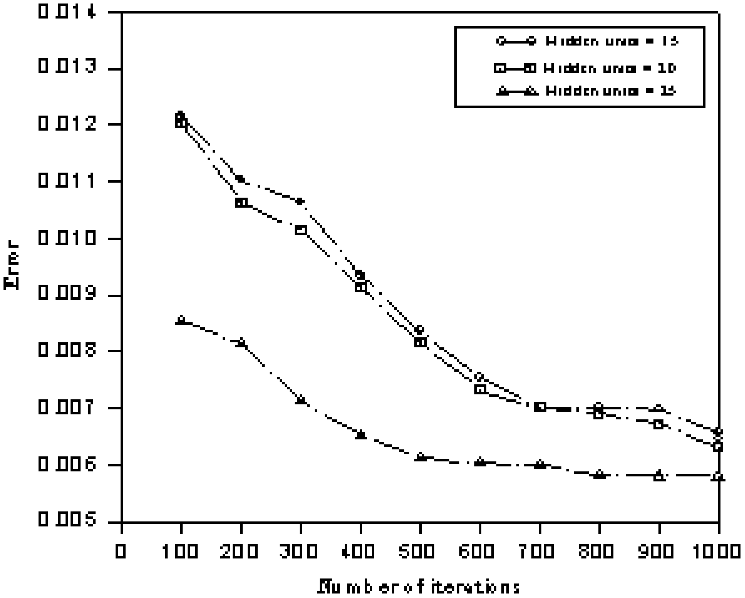
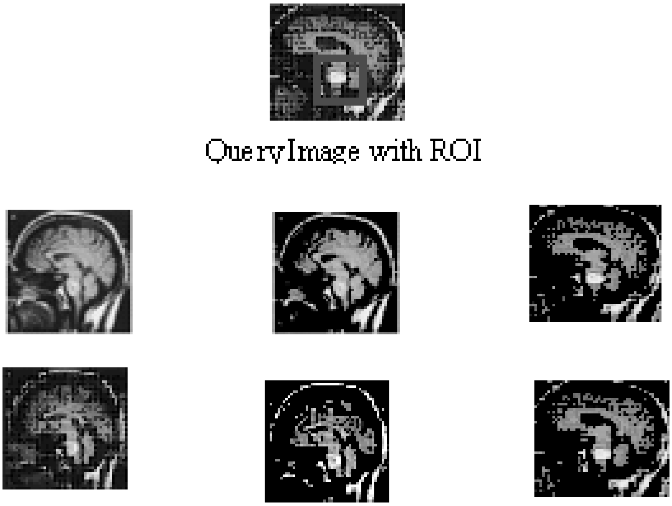


Fig. 2. Learning Profile for Neural Network

Mutation is also a genetic operator, which is to be considered. In the example program, there is a rather high chance of mutation, and then the weights are altered by anything between −1 and 1. The training of neural network is carried out by varying the number of hidden nodes. The learning profile with different number of hidden nodes is shown in Figure 2. The error is reduced with different number of iterations and number of hidden units. The trained output is used for computing the similarity factor, if the value is higher more similar is the image. Thus the Images that have similar regions can be listed out from the image database.

4 Experimental Results

The technique has been implemented in C/C++ and the experiment is carried out on PARAM 10000 parallel computing system with Ultra Sparc II 64-bit RISC CPUs with SUN SOLARIS 5.6 operating system. For testing the effectiveness of the algorithm biomedical image collection from National Technical Information Services Springfield U.S.A., pertaining to tumor has been used as Image Database. The results obtained on applying our algorithms have been consulted with Radiologist from PGI, Chandigarh and the diagnostics matches the medical reports. As such the algorithm has been tested and implemented on about 100 medical images. However, only some of the results are being reported. The diagnostics given by the doctor for Figure 3 is as follows: Brain MRI showed a cavernous malformation located in the pons. Also it is apparent from the Figure the lesion has a different texture and color intensity as marked in the images



QueryImage with ROI

Fig. 3. Learning Profile for Neural Network

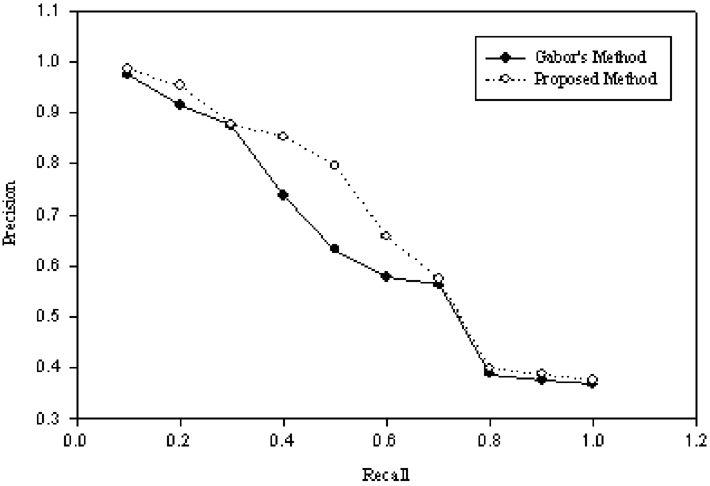


Fig. 4. Comparison of Precision Vs Recall for Gabor method and the proposed method

in Figure 3, the regions are obtained on applying the proposed algorithm. We have compared the retrieval performance of proposed method with the Gabor's Method and the Precision vs. Recall curves are shown in Figure 4.

5 Conclusions and Future Discussions

The study has shown some promise in use of texture and color for extraction of diagnostic information from medical images. The features used characterize different aspects of the texture in a small neighborhood of a pixel in biomedical images. These image features could be used to discriminate among the various tissue types that are inaccessible to human perception. The results obtained from the proposed technique are analyzed, compared and consulted with the Radiologist. The preliminary results of the approach are found to be useful for medical practitioners. The experimental results indicate that this method can be useful for screening biomedical images for any suspected disease that can be diagnosed on the basis of color and texture present in any region of the biomedical image. One major advantage of using neural networks in medical decision support system is that a huge effort of knowledge engineering into the domain knowledge can be saved, provided that sufficient amount of training cases are available. So far only two features of images such as texture and color are only worked upon, as shape also plays a vital role in biomedical image processing, efforts are being made to include the shape feature also.

References

1. Reinus W.R, Wilson A.J, Kalman B, Kwasny S.: Diagnosis of focal bone lesions using neural networks. Investigative radiology 1994; 29(6): 606–611
2. J.R.Smith, S.F.Chang: Local color and texture extraction and spatial query, IEEE Proc. Int. Conf Image Processing, Lausanne, Switzerland, 1996
3. A.Pentland, R.W.Picard, S.Scalroff, Photobook: Content Based Manipulation of Image Databases. IEEE Multimedia pp. 73–75 (1994)
4. J.R.Smith, S.F.Chang: Single color extraction and image query. International conference on Image Processing, Washington D.C., (1995)
5. M.Flickner, H.Sawhney, W.Niblack: Query by Image and Video Content: The QBIC System. IEEE Computer vol.5, pp. 249–274, (1995)
6. C.Faloutsos, R.Barber, M.Flickner, J.Hafner, W.Niblack, D.Petkovic and W.Equitz: Efficient and effective querying by image content. Journal of Intelligent Information Systems, Volume 3, Number 3 & 4, pp. 231–262 (1995)
7. S.Haykin, Neural Networks, Prentice Hall, 1999
8. G.Ciocca, I.Gagliardi, R.Schettini: Retrieving Color Images by Content. Proc. of Image and Video Content based retrieval, (1998)
9. Chad Carson, Serge Belongie, Hayit Greenspan and Jitendra Malik: Region based image querying. Technical Report 97-941, Computer Science Division, University of California at Berkeley, CA, 1997
10. B.M.Mehre, Mohan S.Kakanhalli, A.Desai Narasimhalu, Guo Chang Man: Color Matching for Image Retrieval. Pattern Recognition Letters 16(1995) 325–331
11. W.Y.Ma and B.S.Manjunath: Texture features for browsing and retrieval of image data. IEEE transaction on Pattern Analysis and Machine Intelligence, Volume 18, Number 8, pp. 837–842, August 1996
12. Duda, Hart: Pattern Classification. John Wiley, November 2000
13. R.C.Joshi & Shashikala Tapaswi: Retrieval of Image Databases Using Supervised Learning Approach. CGIV 2002, France, April 2-5, 2002

14. M.E.Peterson, E.Pelikan: Detection of Bone Tumors in Radiographic Images Using Neural Networks. Pattern Analysis and Applications 1999
15. S.Nepal and M.V. Ramakrishna.: Region Identification in CBIR Systems Using Pixel Descriptors. Proc. 4th Intn'l conference on Advances in Pattern Recognition and Digital Techniques, Calcutta, India, Dec27-29, 1999, pp. 215–219
16. S.Kulkarni, B.Verma, P.Sharma, H.Selvaraj: Content Based Image Retrieval using a Neuro fuzzy technique. Proc. Intn'l conference on Neural Networks, July 1999
17. K.Balakrishnan and V.Honavar. Properties of genetic representations of neural architectures. Proce edings of the World Congress on Neural Networks, INNS Press 1995, 807–813
18. J.Branke. Evolutionary algorithms for neural network design and training. Technical Report No. 322, University of Karlsruhe, Institute AIFB, 1995
19. S.Guha, K.Shim et al. CURE : An Efficient Clustering Algorithm for Large Databases

Movie-Based Multimedia Environment for Programming and Algorithms Design

Dmitry Vazhenin, Alexander Vazhenin, and Nikolay Mirenkov

Graduate School Department of Information Systems,
University of Aizu, Aizu-Wakamatsu, 965-8580, Japan,
{d8052102,vazhenin,nikmir}@u-aizu.ac.jp,
http://www.u-aizu.ac.jp/ṽazhenin

Abstract. In the presented work, we introduce a concept of the Movie-Based Programming based on movie-like representation of algorithms and methods. It provides correspondence between algorithmic movie frames and problem solution steps that any frame should visualize/animate a part of a program/algorithm execution. The programming process is in manipulating with special movie-program objects (MP-objects) generating automatically a part of an executable code as well as producing frames, which are adequate to the code generated. It also includes a special multimedia language with high-level constructions and operators in order to make the programming process more efficient and comfortable. Both movie and program can synchronously be generated and debugged. A debugging scheme allows visualizing and controlling all references to the structure elements.

1 Introduction

With the incorporaton of sounds, graphics, animations and video, multimedia is a different way of presenting information and in many cases can enhance traditional programming technologies in order to help users to make a programming process easier and more effective. That is why visual programming technique and languages are widely investigated and used for many applications. They are often involving computer algorithms animation technologies [1] as well as methods for executable code generation from the multimedia specifications [2].

S. Tanimoto proposed the Data Factory, which is an experimental visual programming environment based on a form of computation called the "factory model" [3]. This software supports investigations into the use of the factory model in explaining computing concepts and in exploring the possibilities for programs and program styles in the factory model.

The animated visual 3D programming language SAM (Solid Agents in Motion) for parallel systems specification and animation was proposed in [4]. A SAM program is a set of interacting agents synchronously exchanging messages. The SAM objects can have an abstract and a concrete, solid 3D presentation. While the abstract representation is for programming and debugging, the concrete representation is for animated 3D end user presentations.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 533–541, 2004.

A system called JAVAVIS was developed as a tool to support teaching object-oriented programming concepts with Java [5]. The tool monitors a running Java program and visualizes its behavior with two types of UML diagrams which are de-facto standards for describing the dynamic aspects of a program, namely object and sequence diagrams. We can characterize the most of mentioned systems as very special and focused on solving specific problems.

Multimedia approach for interactive specifications of applied algorithms and data representations is based upon a collection of computational schemes represented in the "film" format proposed in [6,7]. Each scheme by itself reflects some knowledge about a certain method of data processing. When applied to computational methods, a given scheme determines a set of nodes (structure) and/or objects moving in a space-time coordinate system as well as partial order of scanning these nodes and objects. Abstract self-explanatory films, which are series of frames/pictures with different multimedia effects, are used for presentation of the method. Each frame of such a film corresponds to a certain stage of problem solution.

In the presented work, we extend this approach by introducing a concept of the Movie-Based Programming based on movie-like representation of algorithms and methods. It provides correspondence between algorithmic movie frames and problem solution steps that any frame should visualize/animate a part of a program/algorithm execution. The extention is that the programming process is in manipulating with special movie-program objects (MP-objects) generating *automatically* a part of an *executable code* as well as *producing movie frames*, which are *adequate* to the code generated. It also includes a special multimedia language with high-level constructions and operators in order to make the programming process more efficient and comfortable.Both movie and program can synchronously be generated and debugged. A debugging scheme is proposed to allow visualizing and controlling all references to the structure nodes.

In Section 2, we discuss a concept of the Movie-based Programming and show main elements of Movie-based Multimedia Environment for Programming and Algorithms Design (MMEPAD). The third section describes operations on the MMEPAD objects. In Section 4, the movie-program generation process and debugging procedures are shown. The last section contains conclusion and future research topics.

2 MMEPAD Concepts

2.1 Movie-Based Representation of Algorithms and Methods

Usually, a *movie* is a story, play, etc. recorded on a film to be shown in the cinema, television, etc. A *frame* is any of a number (sequence) of small photographs making up a movie (cinema film). We can also consider a computer program or algorithm implementation as a series of computational steps needed to solve a problem. The movie-based programming provides correspondence between frames and solution steps. In this case, any frame should visualize/animate

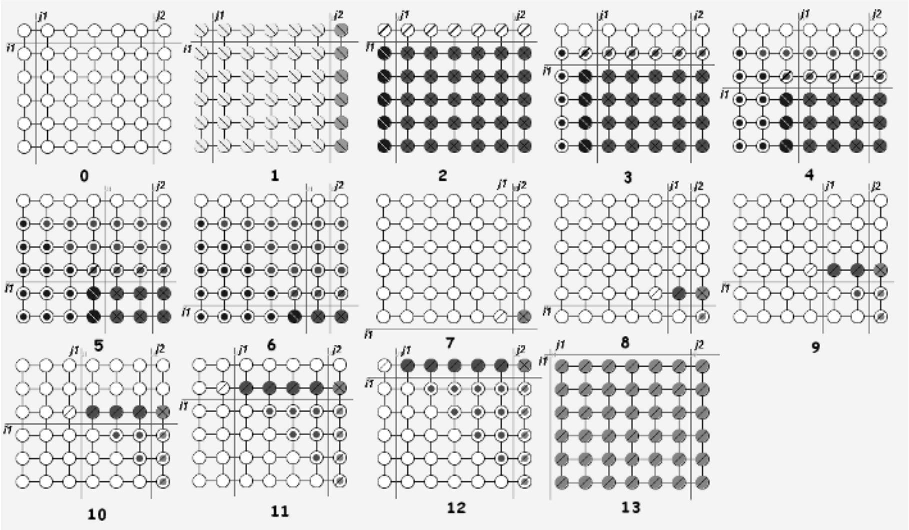


Fig. 1. Gaussian Elimination Algorithm Movie

a part of a program/algorithm execution. We define such a frame as the **Movie-Program Frame** or **MP-frame**.

Figure 1 depicts an example of a movie showing a Gaussian Elimination Algorithm and containing 14 MP-frames. Each MP-frame highlights and flashes some elements of a parameterized matrix. This means that operations or formulas should be defined on a sub-matrix, column, and rows. Different operations can be coded by different colors/sounds/animations. Special **Control Lines** $i1$, $i2$ and $j1$, $j2$ are used to simplify the computational scheme understanding and for possible references. A background can also be used to improve the method representation.

The *Movie-based Programming* is in manipulating with special objects generating a part of an executable code as well as producing MP-frames, which are adequate to the code generated. The key point of proposed concept is a **Movie-Program Skeleton** or **MP-skeleton including** components having these dualistic features (Fig. 2).

The MMEPAD architecture includes five main program modules. The *MP-skeleton Editor* allows users to manipulate with MP-components specifying movie-program parameters. *Movie Generator and Player* is to produce an algorithmic movie by generating a basic MP-frames sequence. This sequence can be extended by additional animations/sounds in order to improve movie presentation. It is also possible to generate and play movie fragments together with editing operations. The *Generator of an Executable Code* is to create a program from the MP-skeleton. There are two possible types of MP-programs. The *Final Executable Program* is generated according to the target machine requirements. The *Movie-based Program* can be implemented and debugged under control of

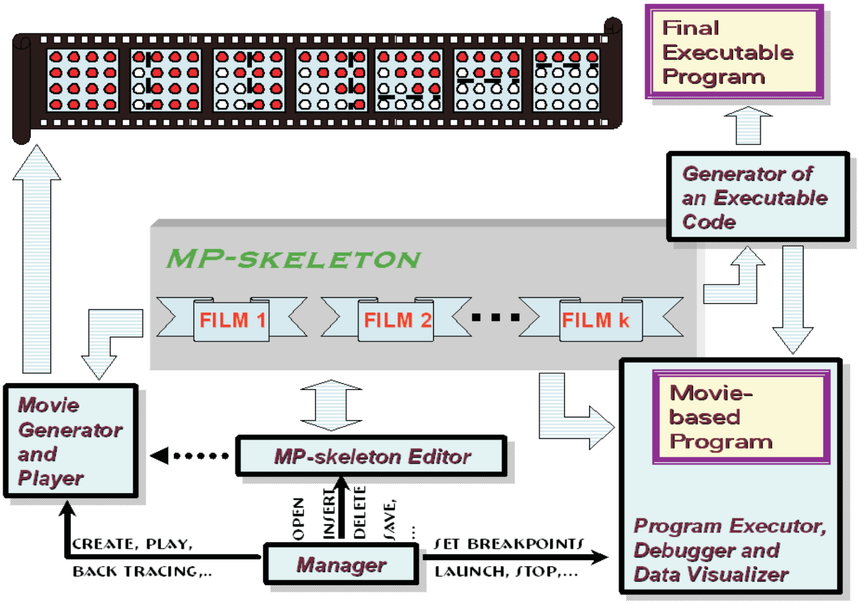


Fig. 2. Main Elements of the MMEPAD System

the *Program Executor, Debugger and Data Visualizer*. The *Manager* controls all system operations and data access procedures.

2.2 MP-Skeleton Components

As shown in Fig. 3, a **MP-skeleton** consists of a set of **MP-films**. Similar to the traditional programming, each film can be considered as a procedure or function. There exists one main MP-film. Other films can be activated by using a special calling mechanism. Each MP-film consists of a set of **MP-stills**. Usually, a still is a photograph of a scene from a movie (cinema film). In the MMEPAD system, such a scene is used for the frames-code generation. The user specifies parameters of this generation by manipulating with the MP-still objects like **MP-nodes**, **MP-structures**, **Control Lines** as well as **MP-formulas** defining operations on these objects.

We can distinguish the following types of MP-stills. A **single MP-still** corresponds to one computational step in the MP-movie. Usually, it produces one MP-frame. A **still series** or **episode** produces the set of MP-frames reflecting a complete fragment/stage of a problem solution. Gaussian Elimination (Fig. 1, Frames 2-6) and Back Substitution (Fig. 1, Frames 7-12) procedures are examples of MP-episodes. The **HEAD-still** should be the first still in any film. It contains description of data structures and variables used in a current film. The **END-still** is to finalize a film.

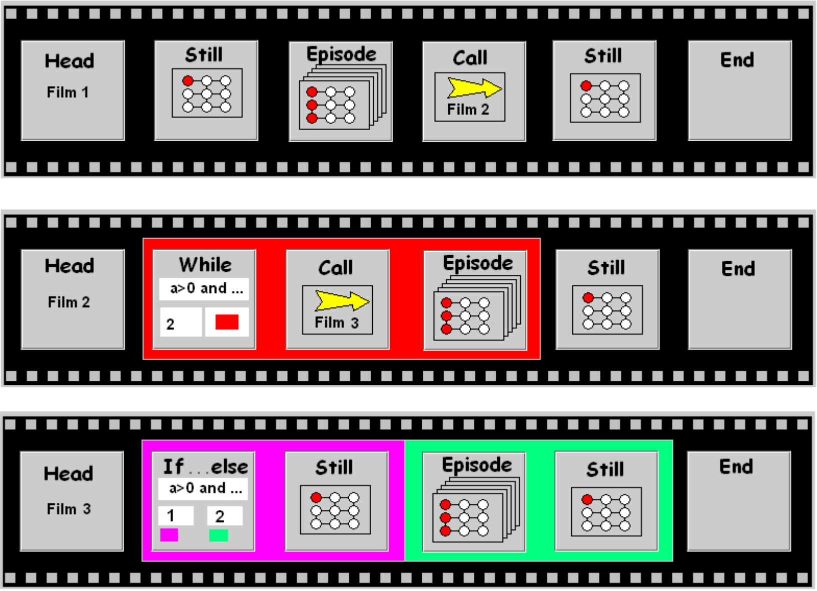


Fig. 3. Example of a MP-skeleton Structure

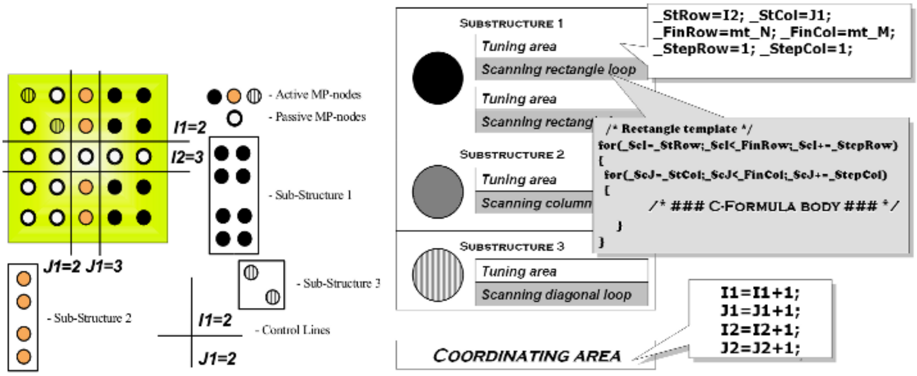
The next group of stills is to control an order of stills processing. **IF-still** is to skip or process selected groups of stills. The user should specify a logical conditional expression as well as mark stills that will be processed for true and false cases correspondingly. The **WHILE-still** is to repeat the processing of stills marked while a condition is true. The **CALL-still** is to pass processing to other MP-films. In this case, the END-still will return control to the parent film.

Importantly, the MP-skeleton structure is very close to a program structure of traditional platforms like C, C++, Java, FORTRAN, etc.

3 Operations on MP-Objects

3.1 Manipulations with MP-Stills

Figure 4a shows an example of MP-still objects. The user can edit a set of these objects as well as specify their parameters. A **MP-node** is an elementary solid multimedia part (cell) of a 3D-space. MP-nodes activities are represented by multimedia attributes (colors, sounds, animations, etc.) A **MP-structure** is a set of MP-nodes joined according to the structure type like scalars, linear and matrix arrays, grids, trees, etc. In Fig. 4, the MP-structure represents matrix data, and each MP-node corresponds to a single matrix element. **Control lines/structures** are used to address nodes, and/or show dependences between nodes.



```
_StRow=I2; _StCol=J1;
_FinRow=mt_N; _FinCol=mt_M;
_StepRow=1; _StepCol=1;
```

```
/* Rectangle template */
for(_ScI=_StRow;_ScI<_FinRow;_ScI+=_StepRow)
{
 for(_ScJ=_StCol;_ScJ<_FinCol;_ScJ+=_StepCol)
 {
        /* ### C-FORMULA BODY ### */
   }
}
```

```
I1=I1+1;
J1=J1+1;
I2=I2+1;
J2=J2+1;
```

a). Main elements

b). MP-templates

Fig. 4. MP-still objects

A **Sub-Structure** is a part/subset of a MP-structure joining equal colored MP-nodes having the same visual/sound representation and implementing the same activities. The configuration of such a Sub-Structure causes partial scanning order of these nodes realized in the **MP-template**. Fig. 4b depicts three typical parts of any MP-template: **tuning area**, **scanning area**, and **coordinating area**. The **tuning area** includes operators/commands setting parameters of scanning loops. These parameters depend on the substructure configuration, placement of control lines or structures, and type of the final program: sequential or parallel. The **Scanning Area** is a set of loops to scan coordinates of MP-nodes in each substructure. The **Coordinating Area** is to provide a correct branch to the next still or complete a MP-frames sequence inside an episode. Figure 5 shows an example of the MMEPAD window for editing MP-stills.

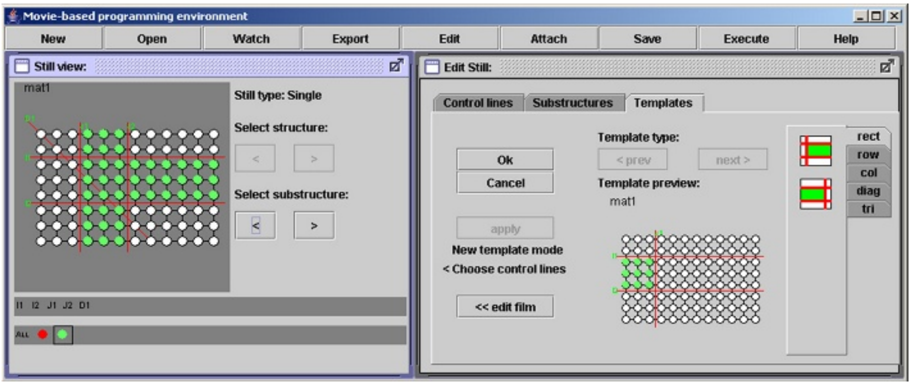


Fig. 5. The MP-still Editor Window

Obviously, the template program implementation depends on MP-structure features. Some difficulties appear when the Sub-structure shape will be *irregular*. We define an *irregular* Sub-structure, which is impossible or difficult to be represented by a partial scanning order of nested loops. To decrease such difficulties, we design a special technique based on decomposing such areas to a set of regular domains having a standard scanning order as shown in Fig. 4b.

3.2 Formulas Attachment

Algorithmic skeleton shows data structures and some activities on these structures. In order to specify these activities, it is necessary to attach arithmetical and/or logical formulas to the skeleton. We distinguish two types of formulas: index (**I-formulas**) and computational (**C-formulas**).

Usually, **index formulas** or **I-formulas** are used to define control lines activities in order to update control lines positions during frame transitions. Each I-formula consists of control line names, basic arithmetical operations and branch conditions. It should be attached to a corresponding control line on a particular still. Therefore, each control line can have several index formulas inside different stills.

Computational formulas or **C-formulas** are necessary to specify operations on active MP-nodes. We define a **C-formula** as a subprogram containing a sequence of arithmetical and logical expressions to specify some local nodes activities. Each **C-formula** includes the following components: **MP-expressions**, **control structures** and **regular text**. **MP-expressions** are to specify data access and operations on MP- nodes. Their notation is very close to the conventional mathematical expressions. Moreover, they may be enhanced by using special multimedia attributes like images, symbols and tables in order to improve the formula perception (Fig. 6). **Control structures** are used to point branch conditions. **Regular text** can be comments and/or a custom code, which extends formula capabilities.

The MP-structure size parameters used in movies may often be different from real ones needed for problem solution. The user should specify not only name and type of a MP-structure but also its dimensions for both movies and program. Those parameters can be either static or dynamic in a program, i.e. may be changed during computations.

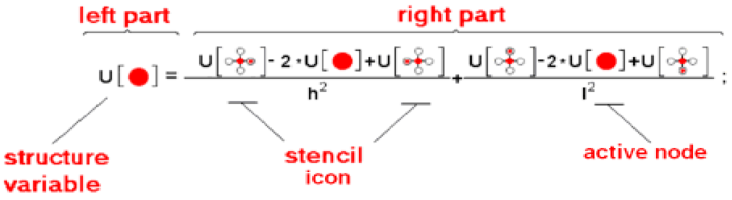
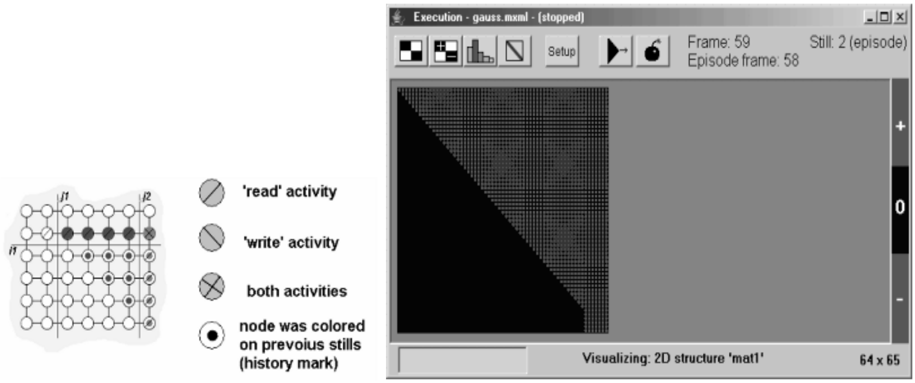


Fig. 6. A C-formula example


j1
j2
i1
'read' activity
'write' activity
both activities
node was colored on prevoius stills (history mark)
Execution - gauss.mxml - (stopped)
Setup
Frame: 59
Episode frame: 58
Still: 2 (episode)
+
0
-
Visualizing: 2D structure 'mat1'
64 x 65

a).MP-formula tracing b). Run-time debugging

Fig. 7. Debugging Environment

4 Movies-Programs Generation and Debugging

Information stored in MP-templates is used to generate MP-frames as well as an executable code. During code generation, some template components (scanning loops, variables, etc.) will be simply transferred in the final code defined by a target system. MP-formulas will be converted to the final code after additional verification. To generate MP-frames, the MP-templates are also used to form images and other graphical information. Calculations using MP-formulas attached can also be implemented, and a movie will be generated representing only one possible case of a MP-program execution obtained according to the real data. It is also possible to generate a movie from a MP-skeleton with non-complete formulas and conditions. In this case, MP-frames with images can only be generated. The user may randomize or specify directly branches needed for IF- and WHILE-stills. As was mentioned, the movie and program have different size parameters of MP-structures. This leads that a movie and program will have/reflect different numbers of MP-frames.

The proposed system allows implementation of a visual debugging of algorithms and programs using two debugging schemes (Fig. 2). The first scheme is to debug film structure and formulas activity during design-time (Fig. 7a). The **MP-formula tracing technique** is used visualizing nodes referred by a formula on a particular frame. Each C-formula is parsed in order to extract indices of nodes where data access is per-formed. Those nodes are marked as active with 'read', 'write' and 'read-write' access type. This allows visualizing any wrong access even before program execution.

The second scheme is to verify movie-based program data-flow during run-time (Fig. 7b) using special breakpoints. When such a breakpoint achieved, the program stops, and the executor invokes the data visualizer. Information provided to the user includes a global frame and still numbers and a frame number inside episode (if any). He/she can choose either to continue/terminate execution, or return to the editing.

5 Conclusion

The proposed concept of the Movie-based Programming allows manipulating MP-objects, *automatically* and *synchronously* generating movie frames and *adequate* executable code for matrix computations. To specify operations on MP-objects, a special multimedia subsystem for the variable declaration and the formula sequence definition was designed. This subsystem uses a special multimedia language with high-level constructions and operators in order to make the programming process more efficient and comfortable. Enhanced text-oriented terms, tables, images and stencils are used for representing the arithmetical and logical expressions.The debugging environment provides additional possibilities to collect and visualize a history of formula references to the structures and data. The MMEPAD system is realized on Java. It generates C/C++ programs and can export movies in the Macromedia Flash Animation format.

Our further works will be oriented in designing movie-based linear algebra library as well as include other structure types like trees, stacks, graphs, etc.

References

1. Stasko, J., Dominique, J., Brown, M., Price, B. (Eds): Software Visualization: Programming As a Multimedia Experience. The MIT Press (1998)
2. Starr. L.: Executable UML. How to build class models. Prentice Hall (2002)
3. Tanimoto, S.: Programming in a Data Factory. In: Proceedings of Human Centric Computing Languages and Environments. Auckland (2003) 100–107
4. Geiger, C., Mueller W., Rosenbach W.: SAM - An Animated 3D Programming Language. In: 1998 IEEE Symposium on Visual Languages. Halifax Canada (1998)
5. Oechsle, R., T. Schmitt, T.: JAVAVIS: Automatic Program Visualization with Object and Sequence Diagrams Using the Java Debug Interface (JDI). In: Software Visualization. Lecture Notes in Computer Science, **2269**, (2002) 1–15
6. Mirenkov N., Vazhenin A., Yoshioka R., et al.: Self-Explanatory Components: A New Programming Paradigm. Int. Jour. of Soft. Eng. and Knowledge Eng. **11** N 1 (2001) 5–36
7. Vazhenin A., Mirenkov N., Vazhenin D.: Multimedia Representation of Matrix Computations and Data. Information Sciences, Elsevier Science Inc. **141** (2002) 97–122

An Improvement Algorithm for Accessing Patterns Through Clustering in Interactive VRML Environments

Damon Shing-Min Liu, Shao-Shin Hung, and Ting-Chia Kuo

Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan 621, Republic of China
{damon,hss,ktc91}@cs.ccu.edu.tw

Abstract. User's traversal paths in VRML environments often can reveal interesting relationships between the objects. However, the massive objects are always stored and scattered in the storage units. This will increase the search time and reduce the system performance. Unfortunately, this problem is never considered in the traditional VRML environments. In this paper, we develop an efficient clustering method to improve the efficiency of accessing objects. The clustering methodology is particularly appropriate for the exploration of interrelationships among objects to reduce the access time. Based on the co-occurrence table and similarity pattern clustering algorithm, we can cluster these patterns more effectively and efficiently. In order to maintain quality of the clusters, the similarity pattern clustering algorithm is presented which satisfies this require-ment. Our experimental evaluation on the VRML data set shows that our algorithm not only significantly cuts down the access time, but also enhances the computational performance.

Keywords: Access Patterns, Clustering, Interactive VRML, Co-Occurrence

1 Introduction

Today, an interactive VRML environment provides virtual navigation with complex 3D models [13] and allows multi-user to traverse in it. Such virtual environment may be a virtual mall or a virtual museum, even a virtual world of an online game. An interactive visualization system can simulate the experience of moving through a three dimensional model, such as a building or an exhibition, by rendering images of the model as seen from a hypothetical observer's viewpoint under interactive control by the user. Several related researches [5] addressed some effective and efficient methods of visibility pre-computing. The models are subdivided into rectangular cells and visibility computations are preformed for those cells. The visibility computations are aimed to find the set of cells visible to an observer able to look in all directions from a position within the cell, and to find the set of objects partially or completely visible to an observer with a

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 542–549, 2004.

specified viewing cone. Nevertheless, they never consider the problem of reducing access times of objects in the storage units. They always concerned about how to display objects in the next frame. In this paper, we consider this problem and solve it by clustering. Clearly, when users traverse in a virtual environment, some potential characteristics will emerge on their traversal paths. Path data should be collected into clusters for users to know their current locations relative to the VRML environments as a whole. The clustering path structure helps users understand the relationships between the objects they have visited. Users can decide where they can go next given their current locations and the objects they have visited so far. By clustering these path structures, it has been made easier for system to predict and pre-fetch the objects for next scene. For example, we can reconstruct the placement order of the objects of 3D model on disk according to common sections of users' path. A new data mining capability for mining the traversal patterns was proposed in [6,7]. They apply data mining technique for mining access patterns in a distributed information providing environment where documents or objects are linked together to facilitate interactive access. Examples for such information providing environments include World-Wide Web (WWW) [8,9] and on-line services, in which seeking for information of interest is realized by traveling from one object to another via the corresponding facilities (i.e., hyperlinks) provided [10]. Similarly, in order to maintain efficiently the massive data objects in virtual environment, we can apply the data mining technique to extract the common features of users' traversal paths, and use the mining results to help us improve the system design in object placement on disk, disk pre-fetch mechanism, or the memory management. Consider the scenario in Figure 1, the rectangles represent objects, and each circle represents a view associated with a certain position. Considering spatial locality, we may take object 1 and object 4 stored onto the same disk block. However, if this view sequence always happens, the mining algorithm will give us different alternative for such situation. The mining algorithm may suggest us to collect object 1, object 3 and object 7 onto the same disk block, instead of object 1 and object 4, due to the temporal coherence.

In our approach, we shall utilize such mining technique to reconstruct our storage organization in constant a period of time. This self-training capability will make our system always be optimized for accessing the objects of large-scale VRML models. Clustering is another main topics in data mining methods [1,3, 4,16]. According to some similarity functions, or other measurements, clustering aims to partition a set of objects into several groups such that "similar" objects are in the same group. It will make similar objects much closer to be accessed together. This results in less access times and much better performance. The quality of clustering has an important effect on predicting the user's traversal behavior. Poor clustering can cause two types of characteristic error: *false negatives*, which are objects that are not accessed, though the user would need them, and *false positives*, which are objects that are accessed, though the user does not need them. In a VRML system, the most essential errors to avoid are false negatives. Because these errors will lead to one or more extra disk access times and thus system performance will be degraded. If we succeed in finding objects which are likely to be used in the near future and cluster them together, that could be a solution to avoid false negative accesses. In this paper, we propose a clustering mechanism based on co-occurrence table and similarity pattern table. Furthermore, a discrimination induction is used to minimize clustering errors by finding desired objects only for users who are likely to use them. To implement the prototype system, a clustering mechanism is also developed. The rest of this paper is organized as follows. Section 2 surveys related works. We define our problem in Section 3. The suggested clustering algorithm is explained in Section 4. Section 5 describes the results of an experimental evaluation. Finally, we summarize our current studies with suggestions for future research in Section 6.

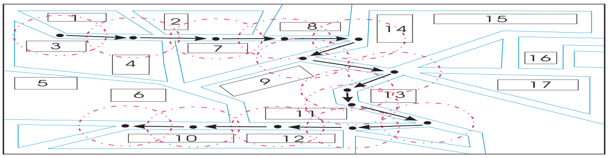


Fig. 1. The circle shows the many objects each view contains and arrow line represents different views when it traverses the path.

2 Related Works

2.1 Spatial Data Structures

In order to manage massive 3D objects of a virtual environment, spatial data structures, such as k-d tree, R-tree, and MD-tree [13], have been applied to some VRML environments. In such systems, only two dimensional shapes which represent the 3D objects such as buildings, electric poles, and so on, are managed instead of the 3D objects themselves. Mainly, this was due to the lack of enough computational power to handle and render 3D objects. In [11], they propose an efficient 3D object management method based on the spatial data structure MD-tree. Using the method, a 3D facility management system that can give rise to the interactive walkthrough of a virtual city is developed. To provide natural view of the scene and the highly interactive environment containing the huge number of objects, they applied the hierarchical spatial data structure, called MD-tree, to perform efficient spatial searches.

2.2 Clustering Methods

Recently, many clustering algorithms have been proposed. In [16], a co-occurrence-based similarity measure for cluster merging was presented. No additional access is needed for evaluating the inter-cluster similarity. However, in our experiments, this algorithm may suggest convergence into one cluster in case that every pattern has somewhat degree of relationship with others. That is, their clustering criteria are not robust for all kinds of data. In [12], to improve the quality of clustering result, they propose the concepts of ”large items” to measure the intra- or inter-similarity of a cluster of transactions. The main idea is that large items should play more important influence in clustering creation and merging phase. Since the ”small item”, presented in [2], contributes to dissimilarity in a cluster. Therefore, they suggest the ”small-large ratio” to perform the clustering. Besides, [2] adopts another direction for clustering by using frequent patterns. They introduce the ”cluster support” and ”global support” to filter the desired frequent patterns. These different support measures can discriminate which cluster the patterns belong to. However, there are several drawbacks in existing clustering methods. First, they only consider access of a single item at a time in the storage units, say hard disks. They only care about how many I/O times the item is accessed. On the other side, we pay more attentions to bulk retrieval, that is, if we can fetch as many objects as possible involved in the same view. This will help to satisfy users' requests more efficiently. Secondly, existing methods do not consider pre-fetch mechanism. Pre-fetch mechanism usually can reduce the I/O seeking times. In other words, most existing methods are forced to seek the desired pattern every time. Finally, when the influence of the disk block size is concerned, how much can we allow to let two or more objects located on different blocks for the purpose of cutting down access times. We will investigate this issue too.

3 Problem Formulation

In this section, we introduce the terms used in our problem and clustering algorithm. Let the problem formulation be a set of m literals called objects (also called items) [7,8]. A view v is denoted by v = $< \chi_1, \chi_2, ..., \chi_k >$, is a unordered list of objects such that each object $X_i \in \Sigma$. The view v is defined as whatever the user stays and observes during the walkthrough of VRML system. A sequence S, denoted by $< v_1, v_2, ..., v_n >$, is an ordered list of n views. Let the database D be a set of sequences (also called transactions). Each sequence records each user's traversal path in VRML system. A sequence $\beta =< \beta_1, \beta_2, ..., \beta_k >$ is a subsequence of sequence $\alpha =< \alpha_1, \alpha_2, ..., \alpha_n >$ if there exists $1 \leq i_1 < i_2 < ... < i_k \leq n$ such that the following $\beta_1 \subseteq \alpha_1, \beta_1 \subseteq \alpha_1, \beta_2 \subseteq \alpha_2, ..., \beta_k \subseteq \alpha_k$ holds. For instance, $< (a)(b,c)(a,d,e) >$ is a subsequence of $< (a,b)(b,c,d)(a,b,d,e,f) >$. But $< (c)(b,e) >$ and $< (a,d)(b)(f,h) >$ are both not subsequences of $< (a,b)(b,c,d)(a,b,d,e,f) >$. Since the former violates the conditions of subsequence: itemsets (c) $\subset$ (a, b) but (b, e) $\subseteq$ (a, b, d, e, f); on the other side, the latter also violates such conditions: (a, d) $\subset$ (a, b), or (f, h) $\subset$ (a, b, d, e, f). The support of a pattern p in the sequence database D is defined as the number of the transactions which contain this pattern p. A frequent pattern is a sequence whose support is equal to or more than the user defined threshold (also called min_support). Let P be a set of all frequent patterns p in D. Finally, we will define our problem as follows. Given a sequence database $D = \{s_1, s_2, ..., s_n\}$, and a set $P = \{p_1, p_2, ..., p_m\}$ of frequent patterns in D. Clustering frequent patterns is the problem of grouping frequent patterns based

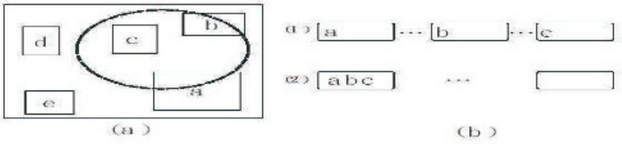


Fig. 2. (left side (a)) means one view in the VRML system. (right side (b)) means two different placements in disks. Upper means the object a, b, and c are dispersed among three disk blocks; lower means the object a, b, and c are combined in the same disk block.

on similarity and consisting of maximizing the intra-cluster similarity while minimizing the inter-cluster similarity.

4 The Co-occurrence and Pattern-Based Clustering Algorithms

In this section, we will analyze the various different problems before we present our clustering algorithm. As seen in Figure 2, consider the right side situation. Upper implies that we will access three disk blocks in order to obtain pattern abc. It spends us three times of access. Lower implies that we only access one disk block for the same purpose. Apparently, lower placement has the advantage of reducing of access times since the co-occurrence patterns are considered. In the viewpoint of clustering, we take another example for demonstration. Given three paths are as follows.
path1:$\{<1,2,3><4,5,6><7,8><18,19,20>\}$;
path2:$\{<1,2,3><9,10><18,19,20>\}$;
path3:$\{<1,2,3><9,10,12><14,15><18,19,20>\}$,
and set min_support=3. After the mining stage, the frequent pattern set is $\{<1,2,3><18,19,20>\}$. But, as for the path3 is concerned, since the buffer is limited and the pre-fetch mechanism is used, $<18,19,20>$ will be filtered out before it was used. This implies that both inter-views and intra-views in paths are required to be considered. To avoid these biased situations, we propose an efficient clustering algorithm that consists of two phases. The first phase is to cluster the co-occurrence patterns on traversal paths as many as possible. In the second phase, for maintaining the high quality of clusters, we introduce the concepts of small-large ratio [15]. Next we will define the data structures and tables used in our algorithm. One table is named Co-Occurrence Table which records the likelihood of any two clusters, say $cluster_i$ and $cluster_j$, co-occurencing in the database D. The Co-Occurrence measure of any $cluster_i$ and $cluster_j$ is defined as $|cluster_i \cap cluster_j| \ / \ |cluster_i \cup cluster_j|$, where $\|$ means the number of elements in the set, $\cap$ means the operation of set intersection and $\cup$ means the operation of the set union. Besides, a pattern is called big if the support exceeds a pre-specified Big-Support. The pattern is called small if the support exceeds a pre-specified Small-Support.

The pattern is called no-thing if the support does not exceed a pre-specified Big-Support, but exceed the Small-Support. SB ratio is a pre-specified value. The Similarity Patterns measures are Intra-Cluster-Cost= $|\cup_{j-1}^{k} Small(C_j)|$, Inter-Cluster-Cost= $\sum_{j-1}^{k} |Big(C_j)|$ - $\cup_{j-1}^{k} |Big(C_j)|$ and Total-Cost(C_j)=Intra-Cluster-Cost(C_j)+Inter-Cluster-Cost(C_j). Finally, the table C is a cluster set which records how many clusters are generated. The Co-Occurrence and Similarity Patterns Clustering algorithm are as follows.

```
Phase I: Co-Occurrence Patterns Clustering Algorithm
// P is the set of frequent patterns. C is the set of clusters, and is set to empty initially.
Input: P, C and Co-Occurrence Table.
Output: C.
Begin
    $C_1$={$C_i$| we set each $pattern_i$ to be a cluster individually, where $pattern_i$ ∈ P}
    $M_1$= Co_Occurence ($C_1$, $\phi$);
    k = 1;
    while | $C_k$ | > n do Begin
        $C_{k+1} = MergeCluster(C_k, M_k)$;
        $M_{k+1} = Co_{O}ccurence(C_{k+1}, M_k)$;
        k = k +1;
    End;
    return $C_k$;
End;
```

```
Phase II: Similarity Patterns Clustering Algorithm
// P is the set of frequent patterns. C comes from the output of previous stage.
Input: P, C, Small-Support, Big-Support and SB ratio.
Output: C.
Begin
    set n=|C|, where || represents the number of clusters in C.
    for i=1 to n do Begin
        Label each pattern as Big, Small, or No-thing in each cluster $C_i$.
        Select the patters whose Small-Large Ratio is above SB ratio and re-distribute these patterns into another clusters if the new total cost is less than the old total cost.
    End;
    return C;
End;
```

5 Performance Evaluation and Analysis

In this section, we will investigate the effectiveness of the proposed clustering algorithm. First, the size of original data set is approximate 335 MB describing

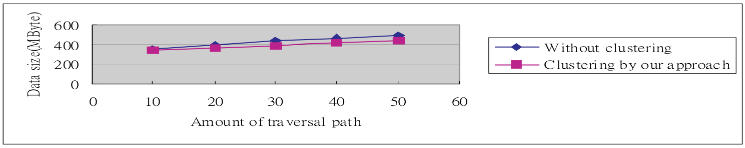


Fig. 3. Comparison of different mechanisms on accessed data size under the different traversal paths.

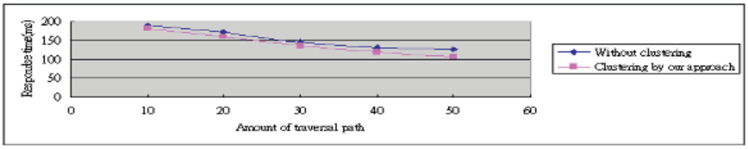


Fig. 4. Comparison of different mechanisms on system response time under the different traversal paths.

1,594 objects in the VRML environments. The traversal paths consist of approximately 10 ∼ 15 views from one end to the other end. In the Figure 3, we can learn that the size of without clustering mechanism is greater than that of with clustering mechanism. The reason is that the net increase (=total size - original size) on clustering is almost half than that of without clustering. This effect influences the performance of Figure 4. Due to the similarity clustering algorithm, we can maintain the quality of clusters as much as possible. Despite the traversal paths increase, with clustering mechanism always gain an advantage over without clustering mechanism. Apparently, not only the access time is cut down but also I/O efficiency is improved. Based on the spatial locality principle, the traditional system without clustering always collects more objects.

6 Conclusions and Future Work

In this paper, a clustering mechanism for VRML system has been presented, which adapts the clustering of spatial objects to the storage system. The proposed clustering mechanism can maintain the quality of cluster. The aim of the this mechanism is to reduce the I/O-cost of the accessing objects. For the VRML database, it can be expected that less additional disk accesses are achieved With properties of co-occurrence similarity table and the big-small ratio clustering scheme are added, it is more precise for us to discover and maintain the frequent traversal patterns. We have been conducting different experiments on datasets so as to find the relationships between extra-views and inter-views. Besides, we also consider how to efficiently cluster the necessary patterns in order to speed up the computations in the future.

References

[1] Diansheng Guo, Donna Peuquet, Mark Gahegan: Opening the Black Box: Interactive Hierarchical Clustering for Multivariate Spatial Patterns. Proceedings of the tenth ACM international symposium on Advances in geographic information systems. (2002) 131–136

[2] Benjamin C.M., Fung, Ke Wang, and Martin Ester: Large Hierarchical Document Clustering Using Frequent Itemsets. Proceeding SIAM International Conference on Data Mining 2003 (SDM '2003). (2003)

[3] L.H. Ungar, D. P. Foster: Clustering Methods For Collaborative Filtering. Proceedings of the Workshop on Recommendation Systems. (1998)

[4] Tian Zhang, Raghu Ramakrishnan, Miron Livny: BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Record , Proceedings of the 1996 ACM SIGMOD international conference on Management of data. (1996) 103–114

[5] Daniel G. Aliaga, Anselmo Lastra: Automatic Image Placement to Provide a Guaranteed Frame Rate. Proceedings of the 26th annual conference on Computer Graphics and Interactive Techniques. (1999) 307–316

[6] Ming-Syan Chen, Jong Soo Park, and Philip S. Yu: Efficient Data Mining for Path Traversal Patterns. IEEE Transactions on Knowledge and Data Engineering. (1998) 209–221

[7] Ming-Syan Chen, Jong Soo Park, and Philip S. Yu: Efficient Data Mining for Path Traversal Patterns. Proceedings of the 16th International Conference on Distributed Computing Systems. (1996) 385–392

[8] Magdalini Eirinaki, Michalis Vazirgiannis: Web Mining for Web Personalization. ACM Transactions on Internet Technology (TOIT). (2003) 1–27

[9] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan: Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorations Newsletter. (2000) 12–23

[10] Mathias Gery, Hatem Haddad: Evaluation of Web Usage Mining Approaches for User's Next Request Prediction. Proceedings of the fifth ACM international workshop on Web information and data management. (2003) 74–81

[11] Y. Nakamura and S. ABE, Y. Ohsawa, and M. Sakauchi: A Balanced Hirtrchical Data Structure Multidimensional Dada with Efficient Dynamic Characteristic. IEEE Transactions on Knowledge and Data Engineering. (1993) 682–694

[12] Ke Wang, Chu Xu, and Bing Liu: Clustering Transactions Using Large Items. ACM CIKM International Conference on Information and Knowledge Management. (1999) 483–490

[13] Y. Nakamura and T. Tamada: An Efficient 3D Object Management and Interactive Walkthrough for the 3D Facility Management System. Proc. IECON'94. (1994) 1937–1941

[14] Eui-Hong Han and George Karypis and Vipin Kumar and Bamshad Mobasher: Clustering Based on Association Rules Hypergraphs. Proc. Workshop on Research Issues on Data Mining and Knowledge Discovery. (1997)

[15] Ching-Huang Yun, Kun-Ta Chuang and Ming-Syan Chen: An Efficient Clustering Algorithm for Market Basket Data Based on Small-Large Ratios. Proceedings of the 25th International Computer Software and Applications Conference (COMPSAC 2001). (2001) 505–510

[16] T. Morzy, M. Wojciechowski, M. Zakrzewicz: Pattern-Oriented Hierarchical Clustering. Proc. of the 3rd East European Conference on Advances in Databases and Information Systems (ADBIS'99).

MPEG-4 Video Retrieval Using Video-Objects and Edge Potential Functions

Minh-Son Dao, Francesco G.B. DeNatale, and Andrea Massa

DIT - University of Trento
Via Sommarive, 14 - 38050 Trento, Italy
{dao,denatale,massa}@dit.unitn.it

Abstract. A novel video retrieval tool based on MPEG-4 video object (VO) representation is presented. The algorithm extends the concept of edge potential functions (EPF), already used in shape-based image retrieval, tailored to work on shapes extracted from VO planes defined in MPEG-4 syntax. First, key frames are selected from the VO sequence by detecting significant object deformations. Then, object boundaries are extracted from each key frame by a simple manipulation of the relevant object plane bitmap, and normalized. Finally a shape-EPF (S-EPF) is calculated from the normalized boundaries and used to perform the matching with user's query. Experimental results demonstrate that the proposed algorithm is efficient and fast in indexing a video sequence according to the presence of specific video objects.

1 Introduction

Everyday, huge amounts of information are collected, produced and communicated in the world. Thanks to increasing bandwidth, better communication facilities and decreasing costs for information storage and processing, users can deal with multimedia information in a broad way. Nevertheless, the problem of easily seeking the requested information is still not completely solved. Today retrieval tools are largely based on textual queries, which are not powerful and flexible enough with respect to the increased amount and variety of available data. Video retrieval is particularly difficult, due to the large amount of information to be processed. In this framework, content-based retrieval is gaining attention as a promising approach to efficiently browse information.

Several content-based retrieval methods have been investigated so far, using low-level features such as color, texture, motion and shape (see [1,2] for a comprehensive survey on the subject). Among different techniques to perform a visual query, the possibility of using object shapes is particularly interesting, for it allows a user to search for a specific visual object by sketching the relevant shape on a simple graphical interface or using a clipboard. On the other hand, this application is very demanding in terms of implementation constraints. In fact, browsing tools should be characterized by a nearly real-time response, low processing requirements, and high robustness. Furthermore, a generic shape-based retrieval scheme working in every situation can be quite difficult to implement

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 550–557, 2004.

due to the presence of complex or cluttered environments, occlusions problems, deformations, etc.

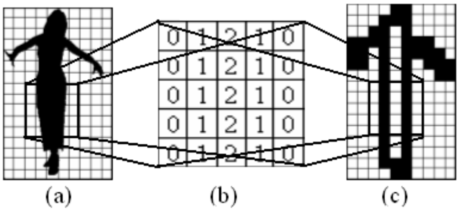
MPEG-4 coding standard [3,4] offers a compressed domain representation of VO shape information, supporting the encoding of multiple VO planes as images of arbitrary shape. Furthermore, it makes very easy to extract the object shape directly from bit stream. Thanks to this simplification, recently a number of systems were proposed using objects as the main key to perform video retrieval in compressed domain. As an example, the Netra-V system [5] provides a VO-based representation for video browsing. In this method, a low-level content description scheme is proposed that uses a new automatic spatio-temporal segmentation algorithm using three visual features: color, texture and motion. The segmented regions are then tracked throughout the video sequence using the relevant local features. The results is a sequence of coherent regions called sub-objects, which are the basic elements for low-level content description. In [6,7] the authors use the VO as a basic unit to build an efficient index for video retrieval in MPEG-4 domain. In their method, birth and death frames of each individual object are found, as well as global motion and camera operations. In [8] Erol and Kossentini proposed a method to easily extract VO shapes from an MPEG-4 compressed stream, and defined a shape similarity measure taking into account the representative temporal instances of each VO together with a set of deformation features such as compactness, eccentricity, Fourier and Angular Radial descriptors.

In this paper, an innovative approach to perform object-based video retrieval is proposed, which is based on the concept of Edge Potential Functions (EPF). First introduced in [9], EPFs mimic the attraction field generated by electrical charges to perform the matching among shape contours. In [9] this approach was successfully applied to sketch-based natural image retrieval. In the present work, the shape of each VO is first extracted from the MPEG-4 syntax in an approximated format [10], and used to compute a shape-based EPF (S-EPF). The potential function is then used to perform the matching among user query and key frames extracted from VOs of the target sequence. This process does not require complex feature extraction and matching algorithms, thus being suited for real-time operation.

In Sect. 2, the extraction of the approximated shape is introduced. In Sect. 3, the concept of EPFs is outlined, while in Sect. 4 the procedure for object-based video retrieval using EPFs is described. Finally, in Sect. 5 a selection of experimental results is presented and discussed and the conclusions are drawn.

2 Approximated Video Object Plane

In MPEG-4 video description syntax, individual VOs are coded into separate bit streams. VO planes (VOPs) contain for each temporal instant the description of the relevant VOs in terms of texture, color and shape information. Texture and shape information are independently encoded, thus allowing to easily extract the shape of each VO from the bit stream without the need to decode the entire



0	1	2	1	0
0	1	2	1	0
0	1	2	1	0
0	1	2	1	0
0	1	2	1	0

Fig. 1. Process of approximating video object plane. (a) the original VOP (b) approximated matrix: intra, opaque and transparent blocks (c) approximated shape

object. Erol et al [10] introduced a technique to efficiently perform this operation. In their approach, the binary alpha planes of the VOP shape are taken into account and approximately decoded on a macroblock basis as follows:

- blocks inside the VOP are transmitted as opaque and are associated to a value 2;
- blocks outside the VOP are transmitted as transparent and are associated to a value 0;
- blocks which contain both pixels inside and outside the VOP, are placed at the VOP boundary and are associated to a value 1.

The advantage of this method is to save computation and reduce the effects of segmentation errors and shape information loss caused by MPEG-4 encoding process. This process is illustrated in Fig. 1.

3 Shape Edge Potential Function

Edge Potential Functions (EPF) were first introduced in [9] with application to content-based image retrieval. This section briefly summarizes the main concepts about EPF. EPF is a conceptual model based on the simulation of the electrical behavior of a charged element in the space. Let assume that a point charge Q is placed in an arbitrary position (x,y,z) in the space: the intensity of the field produced by Q is calculated as:

$$V = \frac{Q}{4r\pi\epsilon} \tag{1}$$

where r is the distance between the position of the charge and the position where the potential is measured, and ϵ is the electrical permittivity constant of the environment. The above definition can be easily extended to consider the potential field due to multiple charges by simply summing the single point charge potentials as follows:

$$V = \frac{1}{4\pi\epsilon}\sum_{i=1}^{N}\frac{Q_i}{r_i} \tag{2}$$

where N is the number of point charges and r$_i$ is the distance between the charge Q$_i$ and field measuring position.

In complete analogy with the above behavior, in our model the potential field is generated from the edges contained in the image, in the sense that every ith edge point at coordinates (x$_i$, y$_i$) is assumed to be equivalent to a point charge Q$_i$. The relevant edge potential field becomes a kind of attractor, whose effect on a test object is to pull a corresponding shape towards the position that maximizes the differential potential (i.e., the best overlapping).

In our proposed method only the shape of a VOP is taken into account. In this case, the electrical permittivity constant can be neglected, and the equivalent charge of every edge point can be set to a unit value. With such hypotheses, equation 2 can be simplified in the following Shape-EPF function (S-EPF):

$$SEPF\left(q,\{Q_i\}\right)=\frac{1}{4\pi}\sum_{i=1}^{N}\frac{1}{r_i} \tag{3}$$

where N is the number of edge points and r$_i$ is the distance between the edge point Q$_i$ and an arbitrary point q.

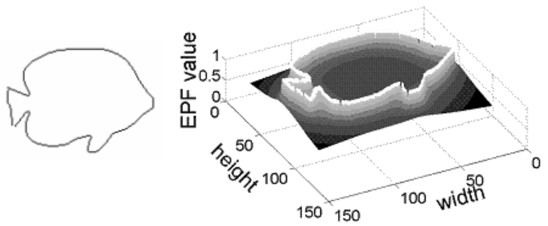
As an example, Fig. 2 shows the S-EPF produced by a well-known MPEG-4 video object test sample.

The further step is to define how S-EPF can be used to match two shapes. To this end, a S-EPF matching function f$_{SEPF}$(A,B) is defined, which measures the similarity of contour shape A with respect to contour shape B, namely:

$$f_{SEPF}\left(A,B\right)=\frac{1}{M}\sum_{i=1}^{M}SEPF\left(a_i,\{b_j\}\right) \tag{4}$$

where a$_i$ and b$_j$ represent the sets of contour pixel of shape A and B, respectively, and M is the number of contour pixels of shape A. Since this function is not commutative, i.e., f$_{SEPF}$(A,B)$\neq$ f$_{SEPF}$(B,A), a possible improvement can be to select as a matching criterion the minimum between f$_{SEPF}$(A,B) and f$_{SEPF}$(B,A), in order to achieve a more reliable shape overlapping measure.

The application of Eq. 4 corresponds to superimposing the query shape to the target S-EPF image and averaging the values of the pixels on S-EPF image



(a) (b)

Fig. 2. Example of S-EPF: (a) shape contour (b) 3D S-EPF

corresponding to the pixels of shape image. It should be noted that when each pixel of the shape image hits the corresponding "charge" (edge) position in the S-EPF image, f$_{SEPF}$ is maximum. When the matching is inexact (e.g., due to deformation), the largest matching measure will anyway occur at a position corresponding to the best overlap.

4 Object-Based Video Retrieval

The proposed method is illustrated in Fig. 3. Here the envisaged application is that a user browses a video server to find videos containing VOs similar to the query one. The process can be split into five parts: extraction of approximated shapes, VOP normalization, computation of S-EPF, key VOP selection, and query processing.

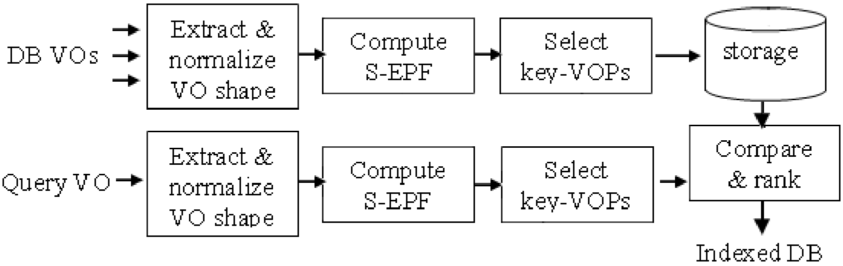


Fig. 3. The VO-based video retrieval process

Extraction of approximated shapes. As mentioned, MPEG-4 allows to separate individual objects in each frame from the stream and extract the relevant boundaries. The contour shape image is constructed accordingly by analyzing the contour blocks as described in Sect. 2.

VOP normalization. The shape is processed to determine the angle β along which it has maximum width, and to rotate the object accordingly. Before rotating, the shape is up-sampled to reduce aliasing problems. Finally, the shape is rescaled to achieve a maximum width equal to a pre-defined value W (typically, W=64) and is aligned to a standard coordinate system to achieve the normalized VOP (see Fig. 4.a-c). During normalization the aspect ratio is kept constant to avoid deformation.

Computation of S-EPF. The S-EPF is computed for each normalized VOP according to Eq. 3 (see Fig. 4.d). The relevant data are attached to the video stream using private data transport structures.

Key VOP selection. The first approximated VOP of the VO, is selected as a key VOP. A new key VOP is selected whenever a significant change occurs in the shape from the previous key VOP. Changes are evaluated by using the matching function in Eq. 4: the heuristic rule adopted is to set up a new key

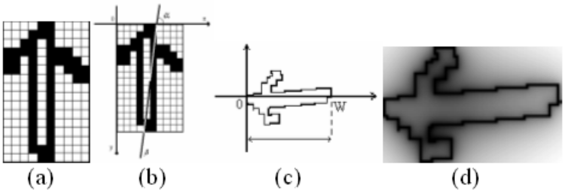


Fig. 4. Shape extraction process: (a) the approximated shape (b) computation of β (c) normalized VOP (d) S-EPF

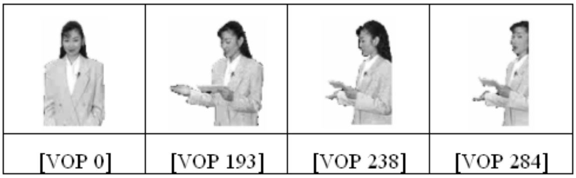


Fig. 5. Key VOP selection of Akyko VO

VOP whenever the value of f$_{SEPF}$ becomes smaller than a threshold γ=0.5. This process continues until the last frame of VO is checked. Fig. 5 show the results of key VOP selection for Akiko sequence.

Query processing. When a user issues a query VO, the query is processed to get the set of normalized S-EPFs associated to the query key VOPs (set A). Then, for each video in the server, the stored S-EPFs associated to the target key VOPs (set B) are scanned and compared to A by using the similarity measure of VOA with respect to VOB which is computed according to Eq. 5.

$$d\left(VO_A, VO_B\right) = \frac{1}{M}\sum_{i=1}^{M} max\left(\left(f_{SEPF}\right)_{j\in[1,N]}\left(VOP_{Ai}, VOP_{Bj}\right)\right) \qquad (5)$$

where M and N are the number of key VOPs belonging to set A and B, respectively. The key VOPs are then ranked according to their increasing distance from the query.

5 Experimental Results

The proposed technique has been extensively tested. We used 20 video streams, each one containing a single video object. The queries have been built as subparts of video object streams. In the selected examples, 3 clips were used (i) a clip of Akiko (CIF) from frame 200 to 250, (ii) the full clip of Irene (QCIF), (iii) the full clip of Foreman (QCIF). The full Akiko and Irene sequences were present also in the test database, while Foreman was not present. Fig. 6a shows the resulting key VOPs.

The result of the three queries on the video base is illustrated in Table 1, where it is possible to observe that in the case of full matching (e.g., full Irene

Table 1. VOs Retrieval Results

Query	**Ranking (3 best and worst)**	**Matching measure (Eq.5)**
Akiko (clip)	Akiko(complete)	0.64
	Irene(complete)	0.57
	Girl(complete)	0.56
	Dancer(complete)	0.2 (worst)
Irene (complete)	Irene(complete)	1
	Girl(complete)	0.62
	Akiko(complete)	0.56
	Fish(complete)	0.25 (worst)
Foreman (clip)	Girl(complete)	0.40
	Irene(complete)	0.37
	Akiko(complete)	0.26
	Dancer(complete)	0.16 (worst)

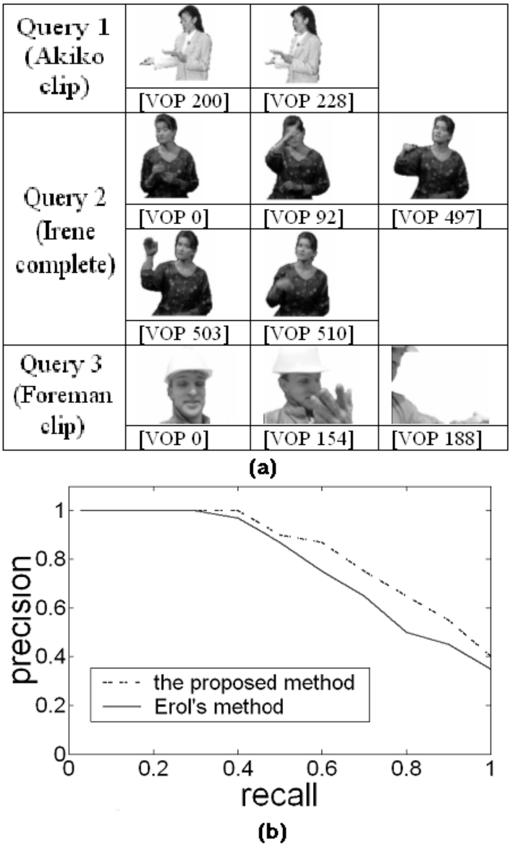


Fig. 6. (a)The query VOPs used for retrieval (b)precision-recall diagram

on full Irene) the matching is maximum, and remains quite high also for partial matching (e.g., subset of Akiko on full Akiko). Vice-versa, in the absence of matching sequence the result is sufficiently low (see, Foreman) to reject the query.

Figure 6b illustrates the precision-recall diagram where our proposed method is compared with Erol's method [8]. Meanwhile Erol uses a lot of shape features as mentioned in Sect. 1, our proposed method uses only information from countour pixels. Therefore, our proposed method dramatically decreases the burden of computation when comparing to Erol's method. Fig. 6b shows that our proposed method is more efficient and effective than Erol's method when retrieving video sequences.

Acknowledgements. This work was partially funded within the framework of the Project PRIN-PER, funded by the Italian Ministry.

References

1. 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, 1999
2. Yoshitaka, A., Ichikawa, T.: A Survey on Content-Based Retrieval for Multimedia Databases. IEEE Trans. on Knowledge & Data Engineering, vol. 11, no. 1, pp. 81–93, 1999
3. ISO/IEC 14496:Coding of audio-visual objects: Video. II edition, 2001
4. Sikora, T.: The MPEG-4 Video Standard Verification Model. IEEE Trans. on Circuits and Systems for Video Technology, vol. 7, no. 1, pp. 19–31, 1997
5. Deng, Y., Manjunath, B.S.: Netra-V: Toward an Object-Based Video Representation. IEEE Trans. on Circuits and Systems for Video Technology, vol. 8, n0. 5, pp. 616–627, 1998
6. Ferman, A.M., Günsel, B. and Tekalp, A.M.: Object-Based Indexing of MPEG-4 Compressed Video. Proc. VCIP'97, vol. SPIE-3024, pp. 953–963, 1997
7. Ferman, A.M., Günsel, B. and Tekalp, A.M.: Motion and Shape Signatures for Object-Based Indexing of MPEG-4 Compressed Video. Proc. ICASSP'97, pp. 2601–2604, 1997
8. Erol, B., Kossentini, F.: Similarity matching of arbitrarily shaped video by still shape features and shape deformations. Proc. ICIP'01, pp. 661–664, 2001
9. S.M. Dao, F.G.B. De Natale, and A. Massa: Edge Potential Functions and Genetic Algorithms for Shape-based Image Retrieval, in Proc. ICIP'03, 2003
10. Erol, B., Kossentini, F.: Automatic Key Video Object Plane Selection Using the Shape Information in the MPEG-4 Compressed Domain. IEEE Trans. in Multimedia, vol.2, no. 2, pp. 129–138, 2000

A Unified Framework Using Spatial Color Descriptor and Motion-Based Post Refinement for Shot Boundary Detection

Wei-Ta Chu[1], Wen-Huang Cheng[2], Sheng-Fang He[1],
Chia-Wei Wang[1], and Ja-Ling Wu[1,2]

[1] Department of Computer Science and Information Engineering
[2] Graduate Institute of Networking and Multimedia
National Taiwan University
{wtchu,wisley,nacci,wjl}@cmlab.csie.ntu.edu.tw, b89075@csie.ntu.edu.tw

Abstract. We propose a unified framework which combines a novel color representation, i.e. spatial color descriptors, and a post-refinement process to detect various types of shot boundaries, including abrupt shot changes, flashlights, dissolves, fade-ins and fade-outs. The spatial color descriptor involving color adjacency and color vector angle histograms incorporates spatial information into color representation and provides robust performance in shot boundary detection. Moreover, a motion-based post-refinement process is developed to effectively eliminate false positives in gradual transition detection, where rapid camera motion or object movement may lead to performance degradation. Experimental results show that these two techniques are integrated seamlessly to give satisfactory performance and present the robustness of spatial color descriptors.

1 Introduction

The development of shot boundary detection algorithms has attracted a large amount of attention in the last decade. There is a rich literature of approaches for detecting video shot boundaries based on color histograms [2], edge pixels [3], motion vectors, and entropy metrics [4]. Although many approaches provide satisfactory results in general cases, few methods are robust to significant appearance changes caused by large-scale object movement or camera motion. One of the solutions to this problem is to design a representation method that takes spatial information into account.

Lee et al. [1] proposed a spatial color descriptor to effectively describe the color distributions and spatial information of video frames. In HLS color space, spatial color descriptors use the metric of color vector angle that is insensitive to variations in intensity, yet sensitive to differences in hue and saturation. When shape or appearance changes, the color pairs at the color edges mostly remain unchanged. Therefore, pixels in a video frame are first classified as either edge or smooth ones and then represented by two color histograms. The proposed color

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 558–565, 2004.

adjacency and color vector angle histograms convey this frame's characteristic. This technique provides robustness to substantial appearance changes and is suitable to be used in image retrieval and video segmentation.

We exploit spatial color descriptors to detect some commonly used shot boundary effects, such as flashlights, abrupt cuts, dissolves, fade-ins and fade-outs. A post-refinement process based on motion analysis is also developed and combined to the framework for eliminating the false alarms caused by rapid camera motion or object movement. This integrated approach is examined by several types of videos and demonstrates its effectiveness on shot boundary detection.

This paper is structured as follows. An overview of spatial color descriptors is stated in Section 2. In Section 3, we describe the proposed framework which hierarchically integrates spatial color descriptors and motion analysis techniques to detect various types of shot boundaries. Section 4 shows the experimental results, and the concluding remarks are given in Section 5.

2 An Overview of Spatial Color Descriptor

Two problems exist in the conventional histogram-based color descriptors. The first one is the lack of spatial information, and the second one is that similar colors are treated as dissimilar because of the uniform quantization of each color axis [1]. To solve these problems, two types of color histograms, i.e. color adjacency histogram for describing edge pixels and color vector angle histogram for describing smooth pixels, are constructed to characterize video frames effectively.

Pixels are classified as edge or smooth pixels based on color vector angle first. A 3 × 3 window is applied to every pixel of a video frame, where the center pixel and neighboring pixels making the maximum color vector angle are used to detect a color edge. If the center pixel in a window is an edge pixel, the global distribution of the color pairs around the edges is represented by a color adjacency histogram based on colors nonuniformly quantized in HLS color space. On the other hand, if the center pixel is a smooth pixel, the color distribution is represented by a color vector angle histogram. The overall distance measure of two successive video frames is represented as

$$d_i = D(i, i+1) = \alpha \times D_{adj}(i, i+1) + \beta \times D_{vec}(i, i+1) \qquad (1)$$

where $D_{adj}(i, i+1)$ and $D_{vec}(i, i+1)$ are distance values (differences of normalized bin values) of color adjacency and color vector angle histograms between frame i and $i+1$, respectively. α and β are scalars for adjusting the weights of two histograms for different genres of videos. In the experiments present in this paper, α and β are both set as 0.5. For video cut detection, if the distance value is larger than a pre-defined threshold, a shot boundary candidate is declared.

3 The Proposed Framework

To robustly address various shot boundary detection issues, we develop a framework which integrates spatial color descriptors and post-refinement techniques,

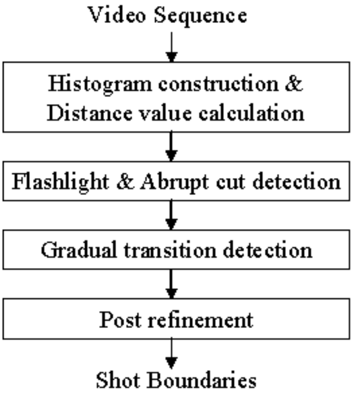

Video Sequence
Histogram construction & Distance value calculation
Flashlight & Abrupt cut detection
Gradual transition detection
Post refinement
Shot Boundaries

Fig. 1. The proposed framework for shot boundary detection

as shown in Fig. 1. After constructing two color histograms, the distance values of successive frames are calculated. To avoid the false alarms caused by flashlights, they are first detected and eliminated before conducting the abrupt cut detection. After detecting abrupt cut, the gradual transition detection process is applied within the range between two abrupt boundaries. Finally, by taking advantage of motion information, a post-refinement process is developed to eliminate some false positives caused by large-scale object movement or camera motion.

3.1 Flashlight and Abrupt Cut Detection

We examine a video sequence by applying a sliding window that spans m frames. The distance values between every two frames, i.e. $d_i, d_{i+1}, ..., d_{i+m-1}$, are used to characterize the behavior of this video segment. If the distance value of two successive frames is larger than a threshold, the frame is declared as a shot boundary candidate. Unfortunately, this simple rule often falsely detects shot boundaries when flashlights occur, which greatly change the luminance of video frames and increase the distance value abruptly.

According to our observation, flashlights often last only one or two frames, and the frames neighboring to a flashlight would have similar color layouts. Therefore, we can detect flashlights by comparing the neighbors of the frame with exploding distance value. The detection rule is defined as follows.

```
if $d_{i+k} > \epsilon$ for $1 \leq k \leq m-2$
   if there exists an $l$, $1 \leq l \leq 4$ such that
      $d' = D(i+k-l, i+k+l) < \epsilon$
      then frame $i+k$ is a flash light
      otherwise frame $i+k$ is an abrupt shot boundary
```

$D(.)$ is the distance value defined in (1) between any two frames, and ϵ is a pre-determined threshold for detecting abrupt discontinuity. In the experiments, the threshold is defined fixedly in the same type of videos without significantly changing the detection performance.

3.2 Gradual Transition Detection

The gradual transitions we considered are dissolves, fade-ins and fade-outs. Unlike abrupt cuts, comparison based on successive frames will not be useful for gradual transition detection because distance values are small during transition [5]. One alternative is to consider local edge information over a series of video frames [3] and match the change patterns of various gradual transitions. However, this method often leads to too many false positives and is not reliable when rapid camera motion or object movement occurs.

In the proposed framework, after detecting abrupt cuts, a gradual transition detection process is applied within the range between two cuts to further explore the structure of this video segment. We exploit the global edge information which is conveyed by the color adjacency histograms. For each frame i, two edge-change values, i.e. edge-increasing value ($E_{i,inc}$) and edge-decreasing value ($E_{i,dec}$), are considered as the metrics for gradual transition detection.

$$E_{i,inc} = \sum_{k=1}^{n} (H_{i,k} - H_{i-1,k}) \quad if \quad H_{i,k} > H_{i-1,k} \tag{2}$$

$$E_{i,dec} = \sum_{k=1}^{n} (H_{i-1,k} - H_{i,k}) \quad otherwise \tag{3}$$

where $H_{i,k}$ is the value of the k-th bin of the color adjacency histogram, and n is the total bin number. Note that different gradual transitions would have different edge change patterns. When a fade-in occurs, the value of edge-increasing will show a peak, while the edge-decreasing value remains smooth. In the case of a dissolve, both edge-increasing and edge-decreasing values would reflect the behavior of edge changes. Fig. 2 shows an example of the curve of edge-increasing values. By using edge change metrics, almost all abrupt cuts have sharp and great-scale peaks. Comparing to the case of abrupt shot change, gradual transitions have smaller change values but are still easily to be distinguished from the frames without shot changes.

An approach based on mean filter is applied to gradual transition detection. Assume that frames i and j $(i < j)$ are declared as abrupt cuts, the rule for detecting fade-ins is defined as follows.

$$\begin{aligned} &E_{mean\_inc} = mean(E_{i,inc}, E_{i+1,inc}, ..., E_{j,inc}) \\ &if \quad (E_{k+l,inc}/E_{mean\_inc}) > \epsilon_g \quad for \quad 1 \le l \le R \quad and \quad i \le k \le j-R \\ &E_{k+l} \quad \text{are frames with fade-ins} \end{aligned} \tag{4}$$

R denotes the width of the sliding window we examined for detecting fade-ins. It is set as 4 in the experiments. ϵ_g is a pre-defined threshold for detecting fade-ins. Similarly, the rules for detecting fade-outs and dissolves are defined through considering edge-decreasing values or the combination of two edge change information.

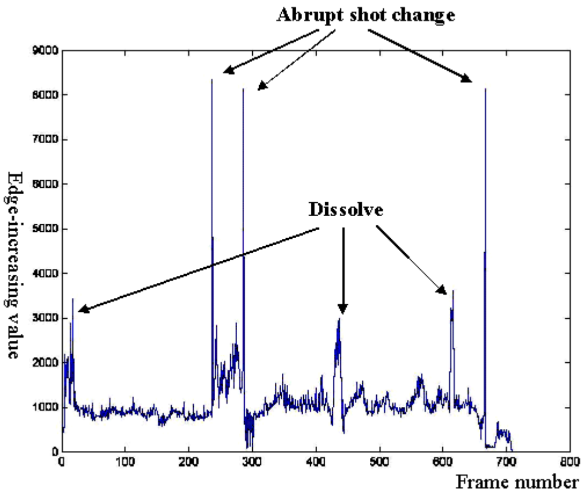


Fig. 2. The curve of edge-increasing values

3.3 Post Refinement by Motion Analysis

Although the spatial color descriptors and edge-based metrics can effectively characterize the behaviors of gradual transitions, some false alarms still exist when extreme camera motion takes place. Therefore, some post-refinement techniques based on motion, priori information, or statistical distributions [7, 8] are proposed to eliminate possible false positives.

In the proposed framework, we adopt a motion-based technique which analyzes motion through spatio-temporal slices processing [9]. The local orientations of temporal slices are estimated by the structure tensor [6]. By modeling the trajectories in tensor histograms, this technique is capable of detecting camera motion so that some false positives can be eliminated.

According to [9], the tensor histograms for horizontal slice with dimension (x, t) and vertical slice with dimension (y, t) are computed. Let $\phi(x, t)_{|y=i}$ and $c(x, t)_{|y=i}$ denote the local orientation and the associated certainty value of a pixel at a horizontal slice in which $y = i$. A 2-D tensor histogram $\mathrm{M}(\phi, \mathrm{t})$ of this video frame in (x, t) dimension is expressed as

$$\mathrm{M}(\hat{\phi}, \mathrm{t}) = \begin{cases} \sum_i \sum_x \sum_t c(x,t)_{|y=i} & \text{if} \quad \phi(x,t)_{|y=i} = \hat{\phi}, \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

which means that each pixel in slices votes for the bin $\phi(x, t)$ with its certainty value $c(c = [0, 1])$. After normalizing by the frame size $m \times n$, the resulting histogram with associated confidence value is represented as

$$C = \frac{1}{T \times m \times n} \sum_{\phi} \sum_{t} \mathrm{M}(\phi, \mathrm{t}), \tag{6}$$

where T is the temporal duration of the video sequence. Detailed descriptions about structure tensor please refer to [9]. In the proposed framework, we detect camera pan and tilt via analyzing the motion in horizontal and vertical slices, respectively. Given a 2-D tensor histogram $\mathrm{M}(\phi, \mathrm{t})$, the tensor orientation ϕ is nonuniformly quantized into three bins, where

$$\Phi_1 = [-90°, -5°),\ \Phi_2 = [-5°, 5°],\ \Phi_3 = (5°, 90°].$$

The scheme quantifies motion information based on its intensity and direction. Φ_1 and Φ_3 represent intense motion, and Φ_2 represents no or slight motion. The normalized 1-D motion histogram N is computed by

$$\mathrm{N}(\Phi_\mathrm{k}) = \frac{\sum_{\phi_i \in \Phi_\mathrm{k}} \sum_t \mathrm{M}(\phi_i, t)}{\sum_{\mathrm{j}=1}^{3} \mathrm{N}(\Phi_\mathrm{j})} \tag{7}$$

Finally, for every three successive frames, they are declared with a camera pan if the following criteria are satisfied in horizontal slices.

$$\begin{array}{l} (\mathrm{N}_{\mathrm{k,k+1,k+2}}(\Phi_1) > \epsilon_N) \wedge (\mathrm{N}_{\mathrm{k,k+1,k+2}}(\Phi_3) < \epsilon_N), \\ (\mathrm{N}_{\mathrm{k,k+1,k+2}}(\Phi_1) < \epsilon_N) \wedge (\mathrm{N}_{\mathrm{k,k+1,k+2}}(\Phi_3) > \epsilon_N), \end{array} \tag{8}$$

where k is the frame index, and ϵ_N is a threshold defined empirically. Similar rules are defined for camera tilt by analyzing vertical slices. Through these processes, the video frames which are declared with both gradual transition and camera motion are discarded from the shot boundary candidates.

4 Experimental Results

We evaluate the proposed framework by using twenty test sequences that belong to four different program categories: news, movies, sports and commercials. They are recorded from TV broadcasts or extracted from MPEG-7 test corpus. Note that these sequences are carefully selected so that they contain many special effects or significant object/camera motions that often cause detection errors. For example, there are many editing effects and dazzling spotlights in selected commercials. The events of camera motion and players walking through screen occur frequently in sports games. The thresholds used in steps described in Section 3 are fixedly defined for different categories of videos without greatly degrading the detection performance. Moreover, to compare the proposed approach with conventional color- and edge-based method, the same test sequences are also applied in the algorithm presented in [3].

Table 1 shows the summary of shot detection results. In general, satisfactory performance could be achieved for different categories of videos. The detection results before and after post-refinement are listed separately to demonstrate the effectiveness of the refinement process. In the gradual transition detection, the refinement process especially shows its effectiveness in sports and commercial sequences because more rapid camera motion and object movement are detected

Table 1. Performance of shot boundary detection

Video Frames	Cut (Correct, False)	Gradual (Correct, False)	Recall/Precsion (Cut)	Recall/Precsion (Gradual)
News (31015)	201(170,1)	25(20,27)	84.57/99.42	48/42.86
-after refinement	201(170,1)	25(12,16)	84.57/99.42	80/42.55
Movie (32123)	312(265,5)	29(11,46)	84.94/98.15	37.93/19.30
-after refinement	312(265,5)	29(10,29)	84.94/98.15	34.48/25.64
Sports (30520)	197(177,10)	37(26,23)	89.84/94.65	70.27/53.06
-after refinement	197(177,10)	37(24,10)	89.84/94.65	64.86/70.59
Commercial	101(90,1)	9(6,4)	89.11/98.9	33.33/50
(5677)	101(90,1)	9(3,3)	89.11/98.9	66.67/60
-after refinement				
Total (with ref.)	811(702,17)	100(49,58)	86.56/97.64	59.32/50.69

Table 2. Comparison of abrupt cut detection between (a) the proposed framework and (b) conventional approach

Video	Recall(a)	Precision(a)	Recall(b)	Precision(b)
News	84.57	99.42	95.6	91.6
Movie	84.94	98.15	99.01	85.71
Sports	89.84	94.65	85.83	39.39
Commercial	89.11	98.9	91.09	92

and eliminated from the shot boundary candidates. Overall, the proposed framework provides 86.56% recall rate and 97.64% precision rate in abrupt cut detection and almost 60% recall and 50% precision rate in gradual transition detection.

Meanwhile, we found that detection accuracy degrades in some cases. Because the spatial color descriptors are based on HLS color space, the color vector angle between two colors with very low or very high intensity would vary significantly even if the Euclidean distance between them is small [1]. That's why the performance of gradual transition in some news and movies sequences is lower than others. This problem can be solved by considering Euclidean distance and color vector angle integrally or slightly modifying the representation of HLS color space.

Table 2 shows the performance comparison between the proposed framework and conventional approach [3]. Only the results of abrupt cut detection are listed because gradual-transition detection was not completely implemented in [3]. Although the conventional approach provides acceptable recall rate in different kinds of videos, it has bad precision performance when there are significant motions in sports video programs. This result shows the reliability of the proposed framework, which takes spatial information and motion-based refinement into account. The proposed approach generally has better precision but worse recall rate. In the current framework, the weights of color vector angle and adjacency histograms are not sedulously adjusted for each video sequence to achieve the best performance. They actually could be assigned by the user to meet different performance requirements, such as higher recall rate with slight degradation in precision.

5 Conclusion

We have presented a framework which integrates spatial color descriptors and motion-based post-refinement techniques to detect various types of shot boundaries. The spatial color descriptors effectively represent the adjacency between colors in video frames and provide robustness to substantial appearance changes. The post-refinement process which exploits structure tensor to detect camera motion is seamlessly combined to improve the detection accuracy of gradual transition. The evaluation results show that this approach provides satisfactory performance in different kinds of videos and is robust to rapid motion and dazzling spotlights. Future work may include improving the performance of motion analysis and conquering the limitation of spatial color descriptors described in Section 4.

References

1. Lee, H.Y., Lee, H.K., and Ha, Y.H.: Spatial Color Descriptor for Image Retrieval and Video Segmentation. IEEE Transactions on Multimedia (2003), Vol. 5, No. 3, 358–367
2. Gargi, U., Kasturi, R., and Strayer, S.H.: Performance Characterization of Video-Shot-Change Detection Methods. IEEE Transactions on Circuits and Systems for Video Technology (2000), Vol. 10, No. 1, 1–13
3. Lienhart, R.: Comparison of Automatic Shot Boundary Detection Algorithms. SPIE Storage and Retrieval for Still Image and Video Databases VII (1999), Vol. 3656, 290–301
4. Cernekova, Z., Nikou, C., and Pitas, I.: Shot Detection in Video Sequences Using Entropy-Based Metrics. Proceedings of International Conference on Image Processing (2002), Vol. 3, 421–424
5. Yeo, B.L. and Liu, B.: Rapid Scene Analysis on Compressed Video. IEEE Transactions on Circuits and Systems for Video Technology (1995), Vol. 5, No. 6, 533–544
6. Granlund, G.H. and Knutsson, H.: Signal Processing for Computer Vision. Norwell, MA: Kluwer (1995)
7. Hanjalic, A.: Shot-Boundary Detection: Unraveled and Resolved? IEEE Transactions on Circuits and Systems for Video Technology (2002), Vol. 12, No. 2, 90–105
8. Lu, H. and Tan, Y.P.: An Effective Post-Refinement Method for Shot Boundary Detection. Proceedings of International Conference on Image Processing (2003), Vol. 2, 1013–1016
9. Ngo, C.W., Pong, T.C., and Zhang, H.J.: Motion Analysis and Segmentation Through Spatio-Temporal Slices Processing. IEEE Transactions on Image Processing (2003), Vol. 12, No. 3, 341–355

HMM-Based Audio Keyword Generation*

Min Xu[1], Ling-Yu Duan[2], Jianfei Cai[1], Liang-Tien Chia[1],
Changsheng Xu[2], and Qi Tian[2]

[1] School of Computer Engineering,
Nanyang Technological University, Singapore, 639798
{mxu,asjfcai,asltchia}@ntu.edu.sg
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613
{lingyu,xucs,tian}@i2r.a-star.edu.sg

Abstract. With the exponential growth in the production creation of multimedia data, there is an increasing need for video semantic analysis. Audio, as a significant part of video, provides important cues to human perception when humans are browsing and understanding video contents. To detect semantic content by useful audio information, we introduce audio keywords which are sets of specific audio sounds related to semantic events. In our previous work, we designed a hierarchical Support Vector Machine (SVM) classifier for audio keyword identification. However, a weakness of our previous work is that audio signals are artificially segmented into 20 ms frames for frame-based SVM identification without any contextual information. In this paper, we propose a classification method based on Hidden Markov Modal (HMM) for audio keyword identification as an improved work instead of using hierarchical SVM classifier. Choosing HMM is motivated by the successful story of HMM in speech recognition. Unlike the frame-based SVM classification followed by major voting, our proposed HMM-based classifiers treat specific sound as a continuous time series data and employ hidden states transition to capture context information. In particular, we study how to find an effective HMM, i.e., determining topology, observation vectors and statistical parameters of HMM. We also compare different HMM structures with different hidden states, and adjust time series data with variable length. Experimental data includes 40 minutes basketball audio which comes from real-time sports games. Experimental results show that, for audio keyword generation, the proposed HMM-based method outperforms the previous hierarchical SVM.

1 Introduction

With the increasing multimedia data available from Internet, there is a need to develop intelligent multimedia indexing and browsing systems. To facilitate high-level abstraction and efficient content-based access, semantics extraction is becoming an important aspect of multimedia-understanding. Recently, video

* This research is partially supported by Singapore A*STAR SERC Grant (032 101 0006).

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 566–574, 2004.

semantic analysis attracts more and more research efforts [1,2,3]. Their works attempt to extract semantic meaning from visual information but little work has been done on the audio parts of multimedia streams.

Audio, which includes voice, music, and various kinds of environmental sounds, is an important type of media, and also a significant part of video. Recently people have begun to realize the importance of effective audio content analysis which provides important cues for semantics. Most of the existing works try to employ audio-visual compensation to solve some problems which can not be successfully solved only by visual analysis [4,5,6,7]. Nepal et al. [5] employed heuristic rules to combine crowd cheer, score display, and change in motion direction for detecting "Goal" segments in basketball videos. Han et al. [6]] used a maximum entropy method to integrate image, audio, and speech cues to detect and classify highlights from baseball video. An event detection scheme based on the integration of visual and auditory modalities was proposed in [4, 7]. To improve the reliability and efficiency in video content analysis, visual and auditory integration methods have been widely researched.

Audio content analysis is the necessary step for visual and auditory integration. Effective audio analysis techniques can provide convincing results. In consideration of computational efficiency, some research efforts have been done for pure audio content analysis [8,9]. Rui et al. [8] presented baseball highlight extraction methods based on excited audio segments detection. Game-specific audio sounds, such as whistling, excited audience sounds and commentator speech, were used to detect soccer events in [9].

In [4,7,9], we used hierarchical Support Vector Machine (SVM) to identify audio keywords. The audio signals were segmented into 20 ms frames for frame-based identification while the audio signals are time continuous series signals rich in context information. By using SVM, we did not take into account the contextual information which is significant for time series classification. HMM is a statistical model of sequential data that has been successfully used in many applications including artificial intelligence, pattern recognition, speech recognition, and modeling of biological sequences [10]. Recently, HMM were introduced to sports video analysis domain [11,12,13,14]. Assfalg et al. [12] used HMM to model different events, where states were used to represent different camera motion patterns. In [14], Xie et al. tried to model the stochastic structures of play and break in soccer game with a set of HMMs in a hierarchical way. Dynamic programming techniques were used to obtain the maximum likelihood play/break segmentation of the soccer video sequence at the symbol-level. These works demonstrated that HMM is an effective and efficient tool to represent time continuous signals and discover structures in video content.

In this paper, we present our recent research work of audio keywords detection by using Hidden Markov Models (HMM) as an improved work for [4,7,9]. In Section 2, we briefly introduce audio keywords and HMM-based generation scheme. Section 3 discusses the audio feature extraction work. Our proposed HMM structure is presented in Section 4. Some comparison experiments and results are listed in Section 5. In Section 6, we draw conclusions and discuss some future work.

Table 1. Audio keywords' relationship to potential events.

Sports	Audio Keywords	Potential Events
Tennis	Applause	Score
	Commentator Speech	At the end (or the beginning) of a point
	Silence	Within a point
	Hitting Ball	Serve, Ace or Return
Soccer	Long-whistling	Start of free kick, penalty kick, or corner kick, Game start or end, offside
	Double-whistling	Foul
	Multi-whistling	Referee reminding
	Excited commentator speech or excited audience sound	Goal or Shot
	Plain commentator speech or plain audience sound	Normal
Basketball	Whistling	Fault
	Ball hitting backboard or basket	Shot
	Excited commentator speech or excited audience sounds	Fast break, Drive or Score
	Plain commentator speech or plain audience sound	Normal

2 Brief Introduction of Audio Keyword Generation System

Audio keywords are defined as some specific audio sounds which have strong hints to interesting events. Especially in sports video, some game-specific audio sounds (e.g. whistling, excited commentator speech, etc.) have strong relationships to the actions of players, referees, commentators and audience. These audio sounds may take place in the presence of interesting events as listed in Table 1. Generally, excited commentator speech and excited audience sounds play important roles in highlight detection of sports video. Other keywords may be specific to a kind of sports game.

Audio signal exhibits the consecutive changes in values over a period of time, where variables may be predicted from earlier values. That is, strong context exists. In consideration of the success of HMM in speech recognition, we propose our HMM based audio keywords generation system. The proposed system includes three stages, which are feature extraction, data preparation and HMM learning, as shown in Fig. 1.

As illustrated in Fig. 1, selected low-level features are firstly extracted from audio streams and tokens are added to create observation vectors. These data are then separated into two sets for training and testing. After that, HMM is trained then reestimated by using dynamic programming. Finally, according to maximum posterior probability, the audio keyword with the largest probability is selected to label the corresponding testing data. We next introduce the proposed system in detail.

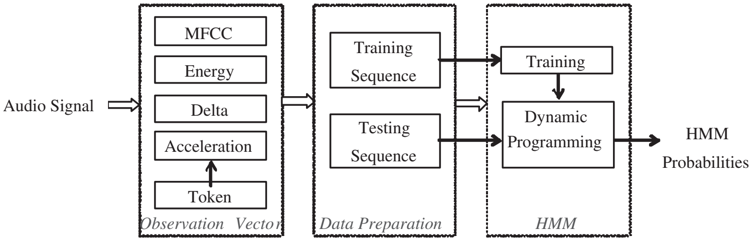


Fig. 1. Proposed audio keywords generation system.

3 Feature Extraction

We segment audio signal at 20 ms per frame which is the basic unit for feature extraction. Mel-Frequency Cepstral Coefficient (MFCC) and Energy are selected as the low-level audio features as they are successfully used in speech recognition and further proved to be efficient for audio keyword generation in [9]. Delta and Acceleration are further used to accentuate signal temporal characters for HMM [15].

3.1 Mel-Frequency Cepstral Coefficient

The mel-frequency cepstrum is highly effective in audio recognition and in modeling the subjective pitch and frequency content of audio signals. Mel scale is calculated as

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}), \tag{1}$$

where $Mel(f)$ is the logarithmic scale of the normal frequency scale f. Mel scale has a constant mel-frequency interval, and covers the frequency range of 0 Hz - 20050 Hz. The Mel-Frequency Cepstral Coefficients (MFCCs) are computed from the FFT power coefficients which are filtered by a triangular band pass filter bank. The filter bank consists of 12 triangular filters. The MFCCs are calculated as

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos[n(k - 0.5)\pi/k], \quad n = 1, 2, \cdots, N \tag{2}$$

where $S_k (k = 1, 2, \cdots K)$ is the output of the filter banks and N is total number of samples in a 20 ms audio unit.

3.2 Energy

The energy measures amplitude variations of the speech signal. The energy is computed as the log of the signal energy, that is, for audio samples $\{s_n, n = 1, \cdots, N\}$

$$E = \log \sum_{n=1}^{N} s_n^2 \tag{3}$$

3.3 Delta and Acceleration

Delta and acceleration effectively increase the state definition by including first and second order memory of past states. The delta and acceleration coefficients are computed using the following simple formula (C_t is the coefficients from feature vector at time t).

$$\Delta(C_t) = C_t - C_{t-1}; Acc(C_t) = \Delta(C_t) - \Delta(C_{t-1}) \tag{4}$$

4 Our Proposed Hidden Markov Model

As for the HMM generation, we need to determine the HMM topology and statistical parameters. In this research, we choose the typical left-right HMM structure, as shown in Figure 2, where $S = \{s_1, \cdots, s_5\}$ are five states; $A = \{a_{ij}\}$ are the state transition probabilities and $B = \{b_i(v_k)\}$ are the observation probability density functions which is represented by a mixture Gaussian density. In our case, each audio frame f_i is regards as one observation o_i. One HMM sample $A = \{f_1, f_2, \cdots, f_n\}$, including n frames, is regards as an observed sequence, $O = \{o_1, o_2, \cdots, o_n\}$. The resulting audio features from each frame form the observation vectors. We use $\lambda = (\Pi, A, B)$ to denote all the parameters, where $\Pi = \{\pi_i\}$ are the initial state probabilities. In training stage, observation vectors are separated into classes to estimate initial B firstly. Then, to maximize the probability of generating an observed sequence, i.e. to find $\lambda^* = arg\ max_\lambda\ p(O|\lambda)$, we use Baum-Welch algorithm to adjust the parameters of model λ.

The recognition stage is shown in Figure 3, where l audio keywords are associated with pre-trained HMMs. For each coming audio sample sequence, the likelihood of every HMM is computed. The audio sequence A is recognized as keyword k, if $P(O|\lambda_k) = max_l\ P(O|\lambda_l)$ [15].

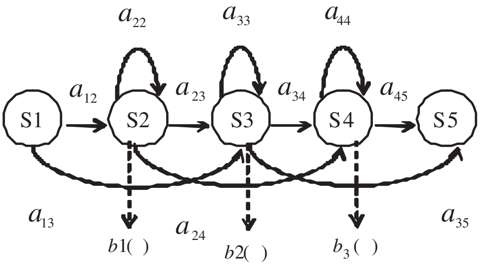


Fig. 2. The Left-right HMM with 5 States.

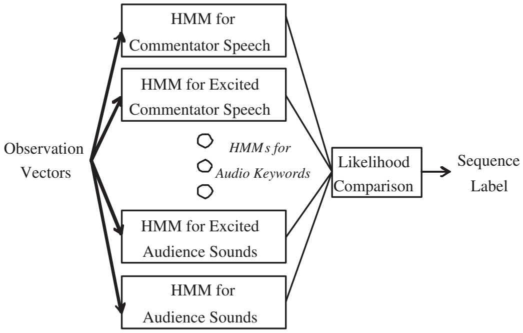


Fig. 3. The HMM overview structure.

In the following experiment, we are concerned about two issues. One is how many states are suitable for a HMM. The other one is the HMM sample length selection.

5 Experiments and Results

Excited commentator speech and excited audience sounds directly correspond to sports highlight which attracts audience's interests mostly. Compared with whistling and hitting ball, the recognition of these two keywords is quite challenging as excited parts always interlace with plain parts. Therefore, in our experiments, we concentrate on excited commentator speech and excited audience sounds.

The audio samples come from a 40 minutes real-time basketball game. They are collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. We used two third for training and one third for testing.

For the HMM learning, different number of states may model different states transition process, which could influence results. Moreover, as each kind of audio keywords have their own durations, we need to choose appropriate sample length for different keyword training. Therefore, we conduct some experiments to compare HMM structures with various states and change HMM sample length to achieve the best performance of our proposed audio keyword generation system.

5.1 HMM with Different Hidden States

At present, we change the states from 3 to 5. The experimental results are listed in Table 2. We find that 3-state HMM is good while 4-state HMM provides better performance for excited commentator. In some sports games, when the environment is very noisy, we can not detect sports highlights only by excited audience sounds while excited commentator speech is able to provide the most important cues. Therefore, higher performance of excited commentator speech identification is necessary. Based on the above criteria and performance results, we thus use 4-states HMM to generate audio keywords.

Table 2. Performance of various HMMs with different states for audio keyword generation.

Audio Keywords	States Number	Recall(%)	Precision (%)
Audience	5 States	95.74	95.74
	4 States	95.74	95.74
	3 States	100	100
Commentator	5 States	100	91.07
	4 States	98.04	94.34
	3 States	100	92.73
Excited Audience	5 States	85.71	85.71
	4 States	85.71	85.71
	3 States	100	100
Excited Commentator	5 States	66.67	100
	4 States	86.67	100
	3 States	73.33	100

Table 3. Performance of different sample length for audio keyword generation (5 states HMM).

Audio Keywords	Sample Length	Recall(%)	Precision (%)
Audience	0.2 Sec	95.39	96.61
	1 Sec	95.74	95.74
Commentator	0.2 Sec	96.52	83.33
	1 Sec	100	91.07
Excited Audience	0.2 Sec	83.33	75.95
	1 Sec	85.71	85.71
Excited Commentator	0.2 Sec	31.65	73.53
	1 Sec	66.67	100

Table 4. Audio keyword generation results (HMM vs. SVM).

Audio Keywords	Methods	Recall(%)	Precision (%)
Whistling	SVM	99.45	99.45
	HMM	100	100
Audience	SVM	83.71	79.52
	HMM	95.74	95.74
Commentator	SVM	79.09	78.27
	HMM	98.04	94.34
Excited Audience	SVM	80.14	81.17
	HMM	85.71	85.71
Excited Commentator	SVM	78.44	82.57
	HMM	86.67	100

5.2 HMM with Different Sample Length

Observation of real sports games reveals that the shortest keyword whistling lasts slightly longer than 0.2 second. Therefore, we segment audio signals into 0.2 second as samples for whistling detection. However, other audio keywords, such as commentator speech, excited audience sounds and etc., last much longer than 0.2 second. Table 3 lists the results of different sample length for several types of audio keywords. The Experimental results show that 1 second sample length is much better than 0.2 second for audience sounds and commentator speech related audio keyword generation. The main reason is that longer sample length provides much more contextual information for HMM to learn in order to differentiate among different audio keywords.

5.3 Comparison Between HMM and SVM

We further do a comparison between the HMM-based method and the SVM-based method [7]. According to the previous experimental results, 4-state left-right structure is selected to build HMM. We choose 0.2 second as sample length for whistling generation and 1 second for other audio keywords (i.e., commentator speech, audience sounds etc.). Compared with SVM-based audio keyword generation, the proposed HMM-based method achieves better performance as listed in Table 4. For the excited keywords generation, which are more significant for highlight detection, the recalls and precisions are improved at least 5%.

6 Conclusion Remarks

Our proposed HMM-based method for audio keyword generation outperforms the previous SVM based method, especially for the excited commentator speech and excited audience sounds. This conforms to the fact that the HMM-based method effectively captures rich contextual information so as to improve different keywords' separability.

As plain/excited commentator speech and plain/excited audience sound are quite general for extensive sports games, we are trying to design adaptive HMMs and combine visual features to boost performance among different kinds of sports games.

References

1. Gong, Y.H., Sin, L.T., Chuan, C.H., Zhang, H.J., Sakauchi, M.: Automatic parsing of TV soccer programs. In: International Conference on Multimedia Computing and System. (1995) 167–174
2. Tan, Y.P., Saur, D.D., Kulkarni, S.R., Ramadge, P.J.: Rapid estimation of camera motion from compressed video with application to video annotation. IEEE Trans. on Circuits and Systems for Video Technology **10** (2000) 133–146

3. Xu, P., Xie, L., Chang, S.F., Divakaran, A., Vetro, A., Sun, H.: Algorithms and systems for segmentation and structure analysis in soccer video. In: IEEE International Conference on Multimedia and Expo. (2001) 22–25
4. Duan, L.Y., Xu, M., Chua, T.S., Tian, Q., Xu, C.S.: A mid-level representation framework for semantic sports video analysis. In: ACM Multimedia. (2003)
5. Nepal, S., Srinivasan, U., Reynolds, G.: Automatic detection of goal segments in basketball videos. In: ACM Multimedia. (2001)
6. Han, M., Hua, W., Xu, W., Gong, Y.H.: An integrated baseball digest system using maximum entropy method. In: ACM Multimedia. (2002) 347–350
7. Xu, M., Duan, L.Y., Xu, C.S., Kankanhalli, M., Tian, Q.: Event detection in basketball video using multiple modalities. In: IEEE Pacific Rim Conference on Multimedia 2003. (2003)
8. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV baseball programs. In: ACM Multimedia. (2000) 105–115
9. Xu, M., Maddage, N.C., Xu, C., Kankanhalli, M., Tian, Q.: Creating audio keywords for event detection in soccer video. In: IEEE International Conference on Multimedia and Expo. (2003) 6–9
10. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
11. Pan, H., Beek, P., Sezan, M.I.: Detection of slow-motion replay segments in sports video for highlights generation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. (2001) 1649–1652
12. Assfalg, J., Bertini, M., Bimbo, A.D., Nunziati, W., Pala, P.: Soccer highlights detection and recognition using HMMs. In: IEEE International Conference on Multi-media and Expo. (2002) 825 – 828
13. Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.S.: Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. (2003) V–632 – V–635
14. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with domain knowledge and hidden markov models. Pattern Recognition Letters **25** (2004) 767–775
15. Young, S., et al: The HTK Book (for HTK Version 3.1) http://htk.eng.cam.edu/. Cambridge University Engineering Department (2002)

Video Scene Segmentation Using Sequential Change Detection

Zhenyan Li[1], Hong Lu[2], and Yap-Peng Tan[1]

[1] Nanyang Technological University, Singapore
[2] Fudan University, Shanghai, P. R. China

Abstract. In content-based video analysis, commonly the first step is to segment a video into independent shots. However, it is rather inefficient to represent video using shot information, as one hour video may contain more than a hundred shots. To address this limitation, most recent work has focused on segmenting a video into scenes, each aggregated by consecutive shots that share similar visual properties or cover a same dramatic event. With the use of sequential change detection and the help of nonparametric density estimation, we propose a novel approach for video scene segmentation in this paper. Experimental results obtained from various test videos suggest that the proposed approach is promising.

1 Introduction

With the advances in media technologies, more and more video data are being generated and forming large databases accessible worldwide. Consequently, efficient methods that can analyze, annotate, search, and retrieve content of interest from large video repositories have become necessary and important.

In content-based video analysis, substantial research efforts have focused on developing techniques to partition a video into shots, each comprising a sequence of consecutive frames that appear to be filmed with a single camera act. However, as one hour video can consist of over a hundred shots and we are mainly concerned with the underlying video stories and subjects, it is rather inefficient and difficult to represent video content based on only shot information.

To address this limitation, more recent work has proposed to group consecutive shots that share similar visual properties or cover a same dramatic event. For example, Kender and Yeo propose a memory-based approach to segment a video into scenes [1]. The approach uses a buffer, which stores the past recent shots, and computes the recall of the incoming shot with the shots in the buffer. If this recall exceeds a pre-determined threshold, a decision is made that the incoming shot covers the same scene as the shots stored in the buffer. The work of [2] proposes a hierarchical scene segmentation approach based on a shot recall measure similar to that in [1]. The approach consists of three steps: initial segmentation, refinement, and adjustment. To speed up the process and maintain segmentation performance similar to that of Kender and Yeo's method, Sundaram and Chang propose a casual, first-in-first-out memory-based model in [3], exploiting both visual and audio information. Approaches that make use of Hidden Markov Model and learning techniques have also been proposed for video scene segmentation [4].

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 575–582, 2004.

In this paper, we propose a novel approach for video scene segmentation by using sequential detection of changes in the video shot sequence. To ensure the proposed approach can adapt well to different video contents, we employ nonparametric density estimation and adaptive thresholding for identifying different scenes. Experimental results obtained from various test videos, as compared with the existing methods, suggest that the proposed approach is rather promising.

2 Sequential Change Detection

Consider each video generating a sequence of shot features accordingly to different feature distributions due to changes of video contents. Our goal is to examine the feature sequence and determine where each video scene change occurs, i.e., the beginning of a new scene. This scene change detection problem can be stated as follows:

> *Given a sequence of M features $\mathbf{X} = \{X_1, X_2, \ldots, X_M\}$ obtained from a video shot sequence, there is an unknown change of the underlying shot feature distribution at k, where $1 < k \leq M$, such that the shot features are generated according to distribution P_1 before k and according to distribution P_2 after (and including) k. The goal is to detect this change (if it exists) while minimizing the false detection rate.*

This problem can be tackled as a testing between a simple hypothesis $\mathbf{H_1}$ (without change) and a composite hypothesis $\mathbf{H}_2 = \cup_{1<k\leq M}\mathbf{H}(k)$ (with change at k), where

$$\begin{array}{lll} \mathbf{H}_1 : & X_n \sim P_1 & \text{for } 1 \leq n \leq M; \\ \mathbf{H}(k) : & X_n \sim P_1 & \text{for } 1 \leq n < k \\ & X_n \sim P_2 & \text{for } k \leq n \leq M. \end{array} \tag{1}$$

With this formulation, the problem can be tackled by considering the log-likelihood ratio — a sufficient statistic for testing between the two hypotheses $\mathbf{H_2}$ and $\mathbf{H_1}$ — defined as:

$$\log \Lambda_1^M(k) = \log \frac{\prod_{n=1}^{k-1} p_1(X_n) \cdot \prod_{n=k}^{M} p_2(X_n)}{\prod_{n=1}^{M} p_1(X_n)} = \sum_{n=k}^{M} \log \frac{p_2(X_n)}{p_1(X_n)}. \tag{2}$$

In particular, the index k, where the change most likely occurs, can be estimated by maximum likelihood estimation $\hat{k} = \arg\max_{1<k\leq M} \log \Lambda_1^M(k)$. Furthermore, if $\log \Lambda_1^M(\hat{k})$ exceeds a pre-defined threshold, a decision can be made in favor of hypothesis $\mathbf{H_2}$; otherwise $\mathbf{H_1}$.

3 The Proposed Approach

The log-likelihood ratio mentioned above can be used only for off-line detection because it can be computed only after all the shot features are available. To detect each video scene change as soon as it occurs and hence to allow the detection of multiple scene changes in the underlying feature distributions, we modify the measure such that a decision of scene change can be made based on the few features around the change location.

In our approach, we select the shot color histogram (SCH) computed in HSV (Hue-Saturation-Value) color space as the feature. Suppose the frame color histogram (FCH) is obtained by uniformly quantizing the HSV color coordinates into 12 (H), 4 (S), and 4 (V) bins, respectively, resulting in a total of 192 quantized colors; a video is segmented into M shots $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$; each shot S_i contains frames $\{f_{b_i}, \ldots, f_{e_i}\}$ with corresponding FCH $\{F_{b_i}, \ldots, F_{e_i}\}$; we compute the bin-wise average color histogram of all the FCHs in the shot as the SCH; that is, $X_i(k) = mean\{F_{b_i}(k), \ldots, F_{e_i}(k)\}$, $1 \leq k \leq B$, where B is the total number of histogram bins.

To modify the measure for detecting scene changes, we consider an important property of the log-likelihood ratio

$$E_{P_1}\left[\log \frac{p_2(X_n)}{p_1(X_n)}\right] \;\leq\; 0 \;\leq\; E_{P_2}\left[\log \frac{p_2(X_n)}{p_1(X_n)}\right] \tag{3}$$

where E_{P_1} and E_{P_2} denote the expectations of the log-likelihood ratio of the shot feature X_n with respect to the distributions P_1 and P_2, respectively. This inequality implies that when the underlying distribution is P_1, the sum of the log-likelihood ratio of shot features has a negative drift and turns to a positive drift after the distribution has changed to P_2. Hence, the existence of a scene change can be decided as soon as the sum of the log-likelihood ratio after turning to a positive drift has exceeded a pre-defined threshold.

It should be noted that the expectation of the log-likelihood ratio after a scene change is commonly used for measuring the distance between two densities, and is known as the Kullback-Leibler divergence [5].

$$D(P_2||P_1) = E_{P_2}\left[\log \frac{p_2(X_n)}{p_1(X_n)}\right]. \tag{4}$$

3.1 Scene Change Measures

For each postulated scene change location k, we measure the Kullback-Leibler divergence using the SCHs in a window of size N, as follows:

$$\begin{aligned} J(k) &= D(P_1||P_2)_{k-N}^{k-1} + D(P_2||P_1)_{k}^{k+N-1} \\ &= \frac{1}{N}\sum_{n=k-N}^{k-1} \log \frac{p_1(X_n)}{p_2(X_n)} + \frac{1}{N}\sum_{n=k}^{k+N-1} \frac{p_2(X_n)}{p_1(X_n)}. \end{aligned} \tag{5}$$

We refer to this as the average-based divergence measure.

Considering that the scene changes of a video can be easily confused by the content changes due to camera and object motion, we also examine another measure, referred to as the median-based divergence measure, by using the median operation instead of the average operation, as follows:

$$J(k) = \text{median}_{n=k-N}^{k-1} \log \frac{p_1(X_n)}{p_2(X_n)} + \text{median}_{n=k}^{k+N-1} \log \frac{p_2(X_n)}{p_1(X_n)}. \tag{6}$$

In our experiments, we set N to 8 for computing these two measures, and set the first N change values to zeros, i.e., $J(k) = 0$, for $k = 1, \ldots, N$.

3.2 Nonparametric Density Estimation

The two measures in (5) and (6) show that the computation of the scene changes requires the knowledge of the two density functions, p_1 and p_2. In the literature, the densities can be approximated by using a parametric density model and estimating the model parameters from representative training data. In video segmentation, however, common parametric models rarely fit well with the feature densities in practice; furthermore, the number of features available for estimating the model parameters is usually limited.

For these reasons, we adopt a nonparametric scheme to estimate the probability densities of each shot feature directly from the neighboring shot features without assuming any models for the underlying distributions. Specifically, we use a k_n-nearest neighbor (with k_n set to 1) to estimate the two corresponding densities of shot feature X_n as follows:

$$\begin{array}{ll} p_1(X_n) = 1 - min_j(\| X_n - X_j \|_q/2), & j = k-N, \ldots, k-1 \\ p_2(X_n) = 1 - min_j(\| X_n - X_j \|_q/2), & j = k+1, \ldots, k+N \end{array} \tag{7}$$

where $\| X_n - X_j \|_q$ is q-norm distance between two shot features X_n and X_j. Here we shall focus mainly on the 1-norm distance for its good performance and ease of computation.

With this nonparametric scheme, we can gauge whether a change in the underlying distribution has taken place. For example, when the shot feature X_k assumes a scene change, the density values estimated for the shot features before X_k with respect to p_1 would be larger than that with respect to p_2, reflecting the fact that the two underlying distributions are different and the shot features before and after X_k follow distributions P_1 and P_2, respectively. On the contrary, when the shot feature X_k does not assume a scene change, the estimated density values with respect to p_1 and p_2 would be similar.

3.3 Scene Change Identification

To identify each scene change as soon as possible, a suitable threshold for change is needed. Although a heuristical threshold can be used for this purpose, it is hard to be well suited for different scene changes due to the large variations of video contents. In our proposed approach, we first identify the peak value of the average-based or median-based divergence measure from a non-overlapping detection window of size $2 \times W + 1$, i.e., the most likely scene change location within the window, and compare it against an adaptive threshold, as follows:

$$\max_{k-W \le i \le k+W} |J(i)| \ge \max(\alpha, \beta z_m) \tag{8}$$

where $z_m = \text{median}_{k-W \le i \le k+W} |J(i)|$, α is a parameter ensuring the scene change is large enough, and β is an appropriate parameter warranting a peak value. In both average-based and median-based divergence measures, we set $2 \times W + 1$ to 11. The two parameters α and β can be determined by using a peer group filtering (PGF) scheme [6] to make the threshold adaptive to each video.

3.4 PGF Based Scheme

The main function of PGF is to separate a set of data into two groups by maximizing the ratio of between-group distance to within-group distance. Suppose $\{d_n, n = 1, \ldots, T\}$ is a distance data sequence ranked in ascending order, where T is the size of the training sequence. The position index l for separating the ranked sequence into two groups can be estimated by maximizing a criterion function C_i, defined as:

$$\hat{l} = \arg\max_i C_i = \frac{|a_{i,1} - a_{i,2}|^2}{s_{i,1} + s_{i,2}}, \quad i = 1, \ldots, T$$

where

$$\begin{array}{ll} a_{i,1} = \frac{1}{i}\sum_{j=1}^{i} d_j, & a_{i,2} = \frac{1}{T-i}\sum_{j=i+1}^{T} d_j, \\ s_{i,1} = \sum_{j=1}^{i} |d_j - a_{i,1}|^2, & s_{i,2} = \sum_{j=i+1}^{T} |d_j - a_{i,2}|^2. \end{array}$$

To estimate parameter α, we apply the PGF scheme to the ranked log-likelihood ratios of a training shot feature sequence $\{L(X_n) = \log\frac{p_2(X_n)}{p_1(X_n)}, n = 1, \ldots, T\}$ in ascending order, and set α equal to the value of the log-likelihood ratio separating the sequence into two groups. To determine parameter β, we form a ratio sequence, $\{R(X_m) = L(X_m)/\text{median}_{j=m-M}^{m-1} L(X_j) + 1, m = M+1, \ldots, T\}$, where $L(X_m)$ is an element in the sequence $\{L(X_n)\}$ and M is set to 5 in our experiments. We rank the sequence $\{R(X_m)\}$ in ascending order and then apply the PGF scheme. The parameter β is set to two times of the value that separates the ranked sequence into two groups. In our experiments, the training sequence is half of the input shot feature sequence, i.e., $T = M/2$.

4 Experimental Results

We have tested the proposed approach to a video set formed by four test videos as shown in Table 1.

Table 1. Scene segmentation performance obtained by the memory-based method [1] and our proposed approach using the average-based (Average) and median-based (Median) divergence measures.

Video	Frame #	Shot #	Scene #	Method	N_c	N_d	Recall	Precision
Tennis	17982	137	5	Memory-based	4	4	0.80	1.00
				Average	5	5	1.00	1.00
				Median	3	3	0.60	1.00
Sitcom	16304	133	10	Memory-based	4	5	0.40	0.80
				Average	7	9	0.70	0.78
				Median	3	3	0.30	1.00
News	28492	187	11	Memory-based	4	8	0.36	0.50
				Average	8	14	0.73	0.57
				Median	6	9	0.55	0.67
Movie	42308	313	22	Memory-based	10	12	0.45	0.83
				Average	14	17	0.64	0.82
				Median	14	16	0.64	0.88

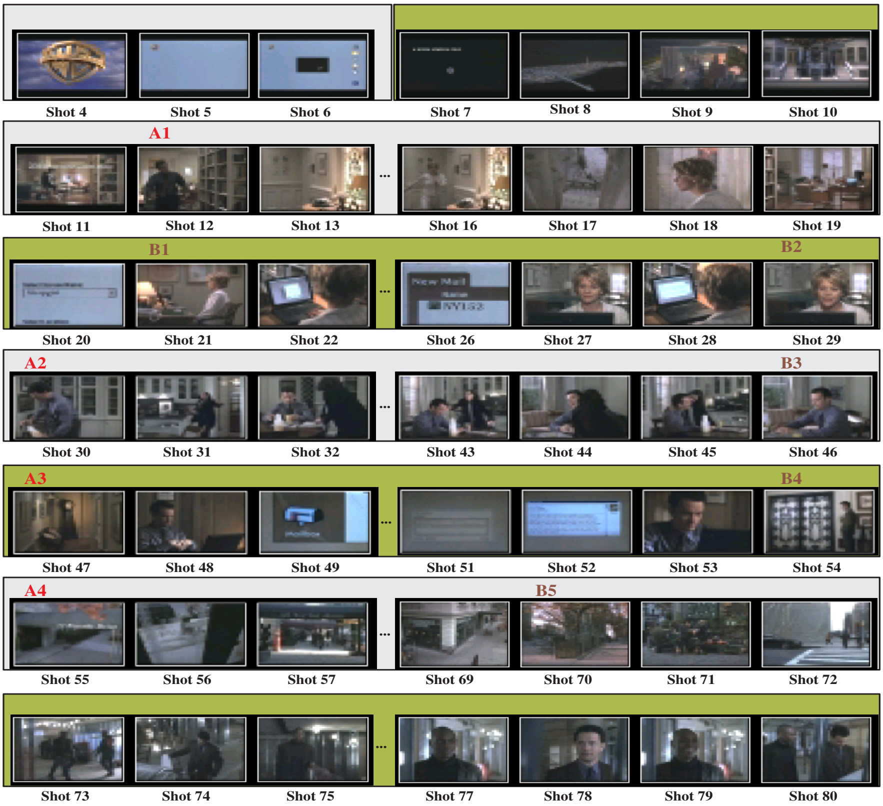


Fig. 1. The ground truth and the scene changes detected by our proposed approach and the memory-based method [1] for the first 80 shots of the test movie video.

Figure 1 shows the scene segmentation results for the first 100 shots of the test movie video obtained by our approach using the median-based divergence measure and the memory-based method proposed by Kender and Yeo [1]. Each ground-truth scene segments of the test video are shown in blocks with different background colors, and each shot is represented by one of its frames. Labels A's are the scene changes detected by our approach, and labels B's are the scene changes detected by the memory-based method.

Figure 1 shows that, although both our proposed approach and the memory-based method can detect most of the scene changes for the test movie video, the location of each scene change detected by our proposed approach is more accurate than that of the memory-based method. For example, three of the four locations of scene changes detected by our proposed approach are the same to that of the ground truth (A2, A3, and A4), and only one scene change location A1 detected by our approach is one-shot different from that of the ground truth. Although these scene changes can be detected by the memory-based method, all of detected the locations (B1, B2, B3, and B4) are one-shot different from that of the ground truth. Furthermore, B5 is a false scene change detected by the memory-based method, and it is avoided by our approach. By comparison, the

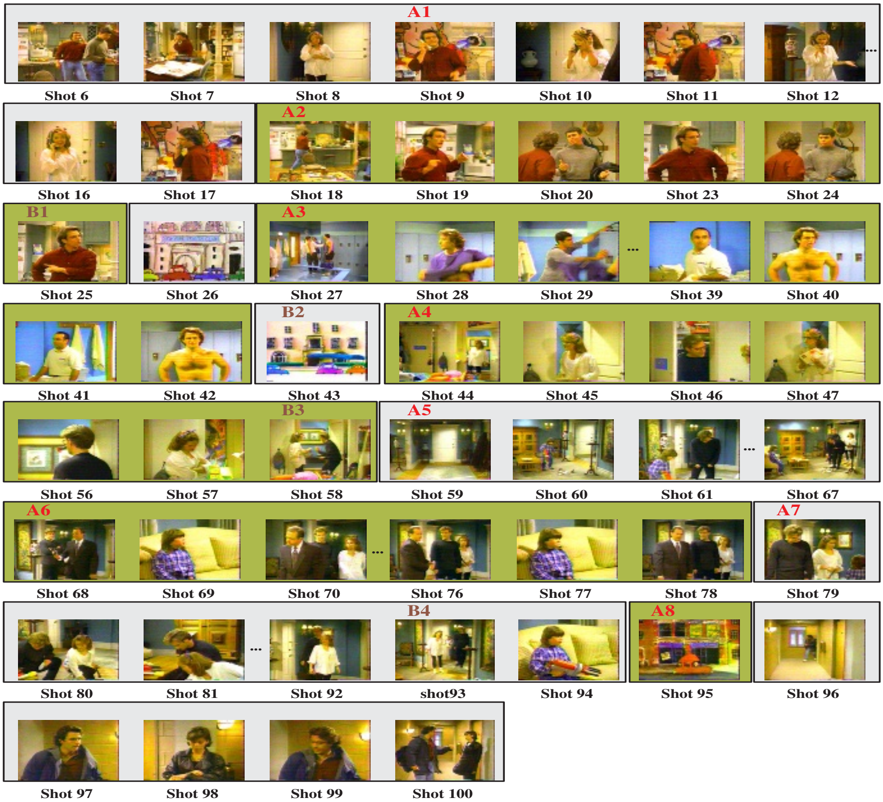


Fig. 2. The ground truth and scene changes detected by our proposed approach and the memory-based method [1] for the first 100 shots of the test sitcom video.

recall and precision for this test movie video obtained by our approach are 0.64 and 0.82 using the average-based divergence measure, and 0.64 and 0.88 using the median-based divergence measure, while the recall and precision obtained by the memory-based method are 0.45 and 0.83, respectively.

Figure 2 shows the scene changes detected by our approach using the average-based divergence measure and the memory-based method for the test sitcom video. It can be observed from the figure that the memory-based method can detect the scene change from an indoor action to commercials (B1 and B2) and the change of indoor scenes (B3). Our approach can detect not only the scene changes detected by the memory-based method (A3, A4 and A5), but also the changes between scenes occurred in a similar place (A2, A6 and A7), which are missed by the memory-based method. Furthermore, A8 is another scene change correctly detected by our approach, while it is missed by the memory-based method. Each location of the scene changes except A1 detected by our approach is the same to that of the ground truth. A1 is a scene change falsely detected by our approach because the first N content changes $J(k)$'s were set to zeros in our experiments. In contrast, the scene changes detected by the memory-based method, B1, B3 and B4, are at least one-shot different from that of the ground truth.

The scene segmentation results obtained by the memory-based method [1] and our proposed approach using the average-based and median-based divergence measures are listed in Table 1 for various test videos. In the table, N_d is the number of scenes detected by each method, and N_c is the number of scenes that are detected correctly. For our approach, the average recall and precision are 0.77 and 0.79 by using the average-based divergence measure, and 0.52 and 0.89 by using the median-based divergence measure. For the memory-based method, the average recall is 0.50 and the average precision is 0.78. The scene segmentation results show that the average segmentation performance of our proposed approach is more accurate than that of the memory-based method. Furthermore, our proposed approach using the average-based divergence measure can achieve a better average detection recall, and that using the median-based divergence measure can normally achieve a higher detection precision.

5 Conclusion

We have proposed in this paper a novel scene segmentation approach using sequential change detection to detect the visual content changes in the underlying shot feature distributions. Two measures based on the Kullback-Leibler divergence, the average-based and median-based divergence measures, are proposed to gauge the changes of visual contents. To identify whether a measured change value signifies a valid scene change, an adaptive threshold selection method based on the peer group filtering is devised to make our approach suitable for various test videos. Experiments on four test video sequences have been conducted and the results suggest that our proposed approach performs better than the existing methods in segmenting a video into scenes.

References

1. Kender, J.R., Yeo, B.-L.: Video scene segmentation via continuous video coherence. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1998) 367–373
2. Kang, H.-B.: A Hierarchical Approach to Scene Segmentation. IEEE Workshop on Content-Based Access of Image and Video Libraries (2001) 65–71
3. Sundaram, H., Chang, S.-F.: Computable scenes and structures in films. IEEE Transactions on Multimedia **4**(4) (2002) 482–491
4. Chaisorn, L., Chua, T.-S., and Lee, C.-H.: The segmentation of news video into story units. IEEE International Conference on Multimedia and Expo, **1** (2002) 73–76
5. Cover, T. M. and Thomas, J. A.: Elements of Information Theory. New York: Wiley (1991)
6. Deng, Y., Kenney, C., Moore, M.S., Manjunath, B.S.: Peer group filtering and perceptual color image quantization. IEEE Intl. Symposium on Circ. and Syst. **4** (1999) 21–24

Feature Extraction and Evaluation Using Edge Histogram Descriptor in MPEG-7

Chee Sun Won

Dept. of Electronic Eng.
Seoul, 100-715, South Korea
cswon@dongguk.edu

Abstract. According to the definition of the edge histogram descriptor (EHD) in MPEG-7, one can easily generate an extra histogram bin from the 5-bin local edge histogram of each 4×4 sub-image. This extra histogram bin defines the ratio of the non-edge area (i.e., monotonous region) in the sub-image. Forming a feature vector with 6 edge/non-edge types, we can generate 33 different feature vectors (or $33 \times 6 = 198$ feature elements) including 16 vectors from 4×4 sub-images, 1 vector from a global histogram, 13 vectors from semi-global histograms, 1 vector from entropy, and 2 vectors from centers of gravity. A statistical hypothesis testing is employed to see which feature vectors/elements are most informative to differentiate different image classes. Experimental results show that non-edge and entropy features are the most informative features among all 33/198 feature vectors/elements.

1 Introduction

Edge in image is an important low-level feature. It can describe both shape and texture features, which are essential elements for content-based image analysis. Its importance makes the edge to be one of the most frequently used image features for the content-based image analysis, demanding a standardized means for its description. In MPEG-7, the edge description in an image is standardized in terms of edge histogram descriptor (EHD) [3][5]. The EHD describes five edge types in the image, namely horizontal, vertical, two diagonal, and non-directional edge types. Population for these five types of edges in a local image region of a sub-image is represented by a histogram with five bins. Specifically, the image space is divided into 16 (4×4) non-overlapping sub-images and for each sub-image a histogram with five edge bins is generated, yielding a combined edge histogram with 80 (16×5) bins for the entire image. The histogram with 80 bins is the standardized edge descriptor for MPEG-7. The apparent role of the EHD is to provide primitive information on the edge distribution in the image. Since the EHD is basically a collection of local edge histograms, it is quite primitive and flexible, allowing us to extract various image features from the 80-bin histogram. For example, in [5], a global and 13 semi-global edge histograms are generated directly from the EHD to be used for the similarity matching. The extended 150-bin (80+14x5) histogram yields better image retrieval performance than that of

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 583–590, 2004.

80 bins only. In addition, according to the definition of the EHD in MPEG-7, one can always generate the sixth feature from the 5 types of edge bins for each 4×4 sub-image. This sixth feature represents the relative population of the non-edge types in the sub-image. Forming a feature vector with 6 edge/non-edge bins, we can extract 33 different feature vectors (or $33 \times 6 = 198$ feature elements) including 16 vectors from 4×4 sub-images, 1 vector from a global histogram, 13 vectors from semi-global histograms, 1 vector from entropy, and 2 vectors from centers of gravity. A statistical hypothesis testing is employed to see which feature elements (or feature vectors), among all 198 feature elements (or 33 feature vectors), are most informative to differentiate different image classes. The purpose of this investigation is to identify descriptor elements that perform well or badly. That is, as a result of this investigation, one can tell which descriptor elements are more informative than others.

2 The Edge Histogram Descriptor (EHD)

Regardless of the size of the given image, the image is first divided into 4x4 sub-images. Each sub-image is a basic region to generate an edge histogram, which consists of 5 bins with vertical, horizontal, 45-degree diagonal, 135-degree diagonal, and non-directional edge types. Since it is required to extract the non-directional edge as well as the four directional ones, a small image block rather than a pixel is needed to extract an edge type [5]. To this end, we further divide the sub-image into non-overlapping image blocks with a small size. Note that the image block may or may not have an edge in it. If there is an edge in the block, we increase the counter of the corresponding edge type by one. Otherwise, the image block has monotonous gray levels and no histogram bin is increased. After examining all image blocks in the sub-image, the 5-bin values are normalized by the *total number of blocks* in the sub-image. Thus the sum of the normalized 5 bins is not necessarily 1. Finally, the normalized bin values are quantized for the binary representation. Since there are 16 (4×4) sub-images, each image yields an edge histogram with a total of 80 (16×5) bins. These normalized and quantized 80 bins constitute the EHD of the MPEG-7.

3 Feature Extraction

Let us denote $b_{ij}(k)$ as a normalized and quantized bin value for a sub-image at $(i, j) \in \Omega$, where $\Omega = \{(i, j); 1 \leq i \leq 4,\ 1 \leq j \leq 4\}$ is a set of indices for sub-images and $k \in \{1, \cdots, 6\}$ indicates one of six edge/non-edge types. More details on labeling (i, j) and k are illustrated in Figure 1. Recalling that the bin values are normalized by the total number of blocks in the sub-image, one can generate the sixth block type (i.e. monotone block) from the five quantized edge bins as $b_{ij}(6) = 1 - \sum_{k=1}^{5} b_{ij}(k)$, where $b_{ij}(6)$ represents the relative population of the number of blocks with the monotonous brightness in the sub-image at $(i, j) \in \Omega$. Note that, as $b_{ij}(6)$ increases, a large monotonous region with no edge is expected in the sub-image at (i, j).

k	Type of the block
1	Vertical Edge
2	Horisontal Edge
3	45-degree Edge
4	135-degree Edge
5	Non-directional Edge
6	Monotone Block

(a)

(1,1) B_1	(1,2) B_2	(1,3) B_3	(1,4) B_4
(2,1) B_5	(2,2) B_6	(2,3) B_7	(2,4) B_8
(3,1) B_9	(3,2) B_{10}	(3,3) B_{11}	(3,4) B_{12}
(4,1) B_{13}	(4,2) B_{14}	(4,3) B_{15}	(4,4) B_{16}

(b)

Fig. 1. Labelling for (a) block types and (b) sub-images with corresponding feature vectors.

3.1 Local Edge Histogram (LEH)

Including $b_{ij}(6)$, we have 6 elements for the histogram of each sub-image at $(i,j) \in \Omega$. Then, we can express the 6 elements as a feature vector $B_l = (b_l(1) \ \cdots b_l(6))^T$ for the l^{th} sub-image at $l = (i,j) \in \Omega$ and $l = 1, \cdots, 16$. Thus, for all 4×4 sub-images, we have a total of 16 feature vectors $\{B_1, \cdots, B_{16}\}$ for local edge/non-edge histograms (LEH).

3.2 Global Edge Histogram (GEH)

For the representation of global edge distribution, we can form the 17^{th} vector B_{17} by averaging the 16 local edge histograms. That is, $B_{17} = (m_1 \ m_2 \ m_3 \ m_4 \ m_5 \ m_6)^T$, where $m_k = \sum_{i=1}^{4} \sum_{j=1}^{4} b_{ij}(k)/16$. The mean vector m_k represents the average occurrence of the edge/non-edge type k in the entire image.

3.3 Semi-global Edge Histogram (SGEH)

In addition to the global edge histogram, 13 semi-global edge histograms (SGEH) are generated from the local edge histograms to represent the edge distribution for specific regions in the image [5]. Note that, for each SGEH, only 4 of the 16 sub-images are chosen. For example, as in Figure 2, we have feature vectors $\{B_{18}, B_{19}, B_{20}, B_{21}\}$ generated by 4 horizontal sub-images and $\{B_{22}, B_{23}, B_{24}, B_{25}\}$ by 4 vertical sub-images. Also, 4 sub-images in the upper-left, upper-right, lower-left, lower-right, and central regions in the image yield 5 additional feature vectors $\{B_{26}, B_{27}, B_{28}, B_{29}, B_{30}\}$, respectively.

3.4 Entropy (ENT)

Since the local histogram bin value $\{b_l(k)\}$ is normalized, it can be treated as a probability mass function. Thus, entropy for the six edge/non-edge features can be obtained from the EHD bins for the 31^{st} feature vector $B_{31} = (H_1 H_2 H_3 H_4 H_5 H_6)^T$, where

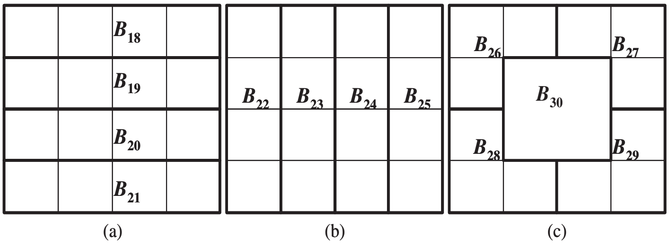


Fig. 2. Feature vectors for semi-global edge histogram descriptors: (a) Horizontal, (b) Vertical, (c) Upper-left, upper-right, lower-left, lower-right, central.

$$H_k = -\sum_{i=1}^{4}\sum_{j=1}^{4} b_{ij}(k) \ln b_{ij}(k). \tag{1}$$

Note that H_k is a measure of randomness and takes small values for smooth or homogeneous texture images.

3.5 Center of Gravity (COG)

In 4×4 sub-images, the center of gravity (COG) of each edge/non-edge type can tell us its density bias in the 4×4 sub-image grid. For the uniformly and symmetrically distributed edge/non-edge type, its COG will be in the center of the grid, which is $(2.5, 2.5)$. However, if the density of an edge/non-edge type for a sub-image is higher than others, then its COG will be closer to the grid of that sub-image. The 2-D components of the center of gravity for the edge/non-edge type k are the last two feature vectors $B_{32} = (c_x(1)\ c_x(2) \cdots\ c_x(6))^T$ and $B_{33} = (c_y(1)\ c_y(2) \cdots\ c_y(6))^T$ defined as follows

$$c_x(k) = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} i \times b_{ij}(k)}{\sum_{i=1}^{4}\sum_{j=1}^{4} b_{ij}(k)}, \tag{2}$$

$$c_y(k) = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} j \times b_{ij}(k)}{\sum_{i=1}^{4}\sum_{j=1}^{4} b_{ij}(k)}. \tag{3}$$

4 Feature Evaluation Using Statistical Hypothesis Testing

Let the value of a feature element $b_l^{(m,n)}(k)$ be the realization of the random variable $B_l^m(k)$, where $l \in \{1, 2, \cdots, 33\}$, $n \in \{1, \cdots, N\}$, m, and k denote the indices for the feature vectors, image sample, image class, and the edge/non-edge types, respectively. So, for an image class m, there are N image samples. Now, our question is whether the features vectors obtained from the sample images of two image classes differ significantly. This is equivalent to asking whether the feature elements in $B_1, \cdots, B_{33}$ are informative enough. A way of evaluating the informativeness is to employ the statistical hypothesis testing [4]. Suppose that there are two image classes ω_1 and ω_2 and, for each random variable $B_l^m(k)$, we have N realizations (or samples) $\{b_l^{(1,1)}(k), \cdots, b_l^{(1,N)}(k)\}$ and $\{b_l^{(2,1)}(k), \cdots, b_l^{(2,N)}(k)\}$ for ω_1 and ω_2, respectively. Then, we will try to answer which of the following hypotheses is correct:

H_1 : The sample values $\{b_l^{(1,1)}(k), \cdots, b_l^{(1,N)}(k)\}$ and $\{b_l^{(2,1)}(k), \cdots, b_l^{(2,N)}(k)\}$ for the k^{th} feature element of the l^{th} feature vector differ significantly.

H_0 : The sample values $\{b_l^{(1,1)}(k), \cdots, b_l^{(1,N)}(k)\}$ and $\{b_l^{(2,1)}(k), \cdots, b_l^{(2,N)}(k)\}$ for the k^{th} feature element of the l^{th} feature vector do not differ significantly.

The decision on the hypotheses H_1 and H_0 is based on the consideration of the differences of the sample mean values. Specifically, we will test whether the difference of the mean values of the two classes is significantly different from value zero. To this end, we need the means $\mu_1 = E[B_l^1(k)]$ and $\mu_2 = E[B_l^2(k)]$ and the variances $\sigma_1^2 = E[(B_l^1(k) - \mu_1)^2]$ and $\sigma_2^2 = E[(B_l^2(k) - \mu_2)^2]$ for the classes ω_1 and ω_2, respectively. (For notational convenience, we drop the indices l and (k) for μ_1, μ_2, σ_1^2, and σ_2^2.) If the means and the variances are not known, the sample means and sample variances are used (i.e., $\hat{\mu}_m = \sum_{n=1}^N b_l^{(m,n)}(k)/N$ and $\hat{\sigma}_m^2 = \sum_{n=1}^N (b_l^{(m,n)}(k) - \hat{\mu}_l)^2/(N-1)$, where $m = 1, 2$). Having the statistics $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$, let us define the test statistic q as follows

$$q = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)\sqrt{\frac{2}{N}}}. \tag{4}$$

Under the assumption that the random variables $B_l^1(k)$ and $B_l^2(k)$ for classes ω_1 and ω_2, respectively, have normal distributions with the same variance, the random variable of q turns out to follow the t-distribution with $2N-2$ degrees of freedom [4].

Let us define the acceptance interval D as the interval for q in which the hypothesis H_0 takes a high probability. That is, if the value of q obtained from (4) lies in D, we will accept H_0. Otherwise, for its complement $\bar{D}$, if q lies in $\bar{D}$, we will reject H_0 and accept H_1. Under the assumption that H_0 is true, the conditional probability denoted ρ that q lies in $\bar{D}$ given the hypothesis H_0 is the probability of an error in our decision. Note that, given ρ, the interval D (also $\bar{D}$) is determined. As the error probability ρ decreases, D covers a wider interval, which requires larger q value to lie outside D (i.e., inside $\bar{D}$) for H_1. This implies that, as $|q|$ increases, the acceptance probability of the hypothesis H_1 remains to be high even for the low error probability ρ. This also implies that $|q|$ can be used as a measure for the ability of differentiating two image classes. That is, a bin (or a feature) with a larger $|q|$ value can be considered to be more informative.

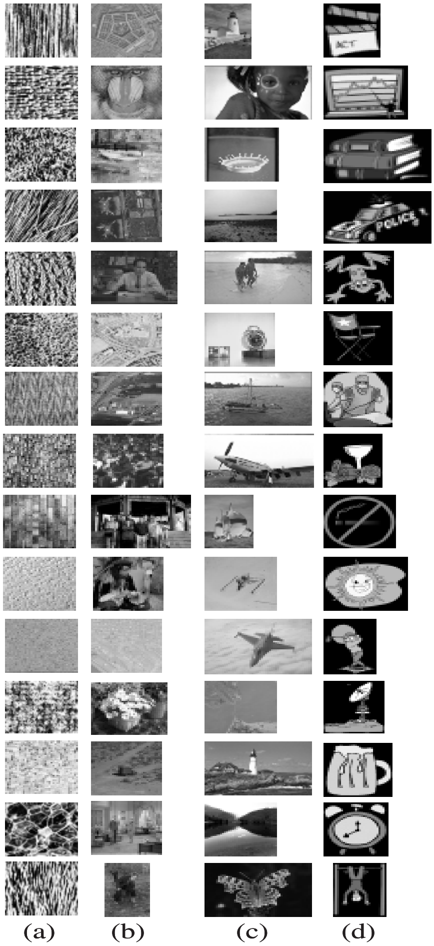


Fig. 3. Image samples for (a) homogeneous texture: image class A, (b) non-homogeneous texture/complex: image class B, (c) image objects with monotonous background: image class C, (d) clip art image: image class D.

5 Experiments

As shown in Figure 3, four image classes with 15 sample images (i.e., $N = 15$) for each class are used for our experiments, namely homogeneous texture images (image class A), non-homogeneous complex images (image class B), and monotone background images (image class C), and clip art images (image class D). The hypothesis testing is conducted for each possible pair of the 4 image classes (i.e., a total of 6 pairs). Note that there are 5 different feature groups, namely LEH with $\{B_1, \cdots, B_{16}\}$, GEH with B_{17}, SGEH with $\{B_{18}, \cdots, B_{30}\}$, ENT with B_{31}, and COG with $\{B_{32}, B_{33}\}$. Among 33 feature vectors, dynamic

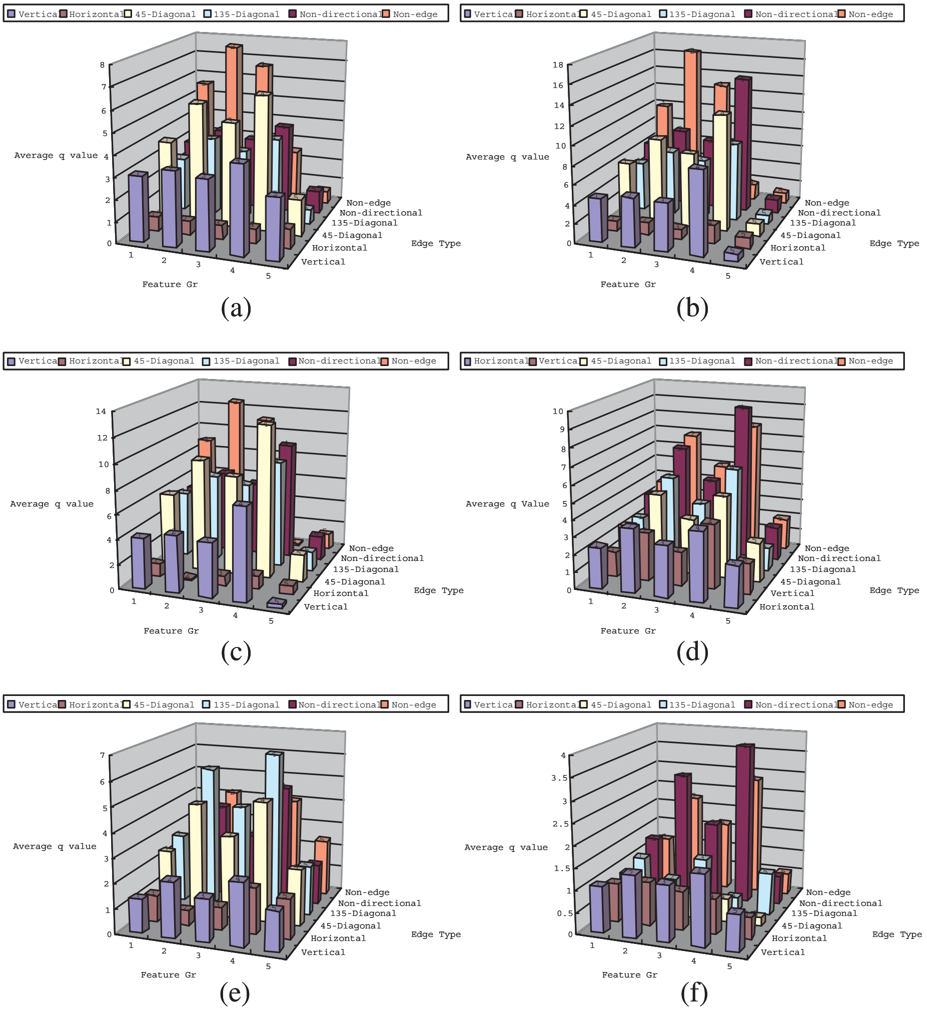


Fig. 4. Average $|q|$ values for feature groups labelled as 1:LEH, 2:GEH, 3:SGEH, 4:ENT, 5:COG between image classes of: (a) A and B, (b) A and C, (c) A and D, (d) B and C, (e) B and D, (f) C and D.

ranges for LEH, GEH, and SGEH are within $[0, 1]$. Thus, for fair comparisons, feature elements of ENT and COG are needed to be normalized. For the normalization of ENT we can divide the entropy in (1) by $16/6 \times ln(1/6)$, where $b_{ij}(k) = 1/6$ yields the highest entropy. Also, the COG can be normalized by dividing (2) and (3) by 5. Since feature vectors within the same feature group exhibit similar statistical properties, the $|q|$ values with the same feature groups are averaged and plotted in Figure 4. As one can see in the figure, feature vectors of GEH, SGEH, and ENT yield large $|q|$ values, demonstrating their excellent informativeness. We also observe that most $|q|$ values are large enough to guarantee low error probabilities. Among all 6 pairs of experiments, however, the $|q|$ values obtained for the classification between image groups C and D are smaller than others, meaning the relative weakness of the edge-based classification for them. For most cases (especially for GEH), the sixth bin $b_l(6)$ takes higher $|q|$ values, showing its high informativeness.

6 Conclusions

This paper evaluates 198 feature elements including 80 bins of EHD in MPEG-7 and other 118 features generated from them. The evaluation is based on the hypothesis testing framework. For a pair of image classes, the statistic $|q|$ is calculated for each feature element including the sixth bin for the non-edge type. Experimental results with our choice of four image classes show that the global and the semi-global histograms and the entropy yield large $|q|$ values for most of cases, meaning low classification error probabilities. In particular, we observed that the non-edge type feature generated from the 5 edge types for each sub-image is indeed most informative in differentiating different images.

Acknowledgements. This work was supported by University IT Research Center Project.

References

1. Bigun, J., Gustavsson, T.: Defect image classification and retrieval with MPEG-7 descriptors, Lecture Note in Computer Science, vol. 2749, Springer-Verlag, (2003) 349–355
2. Eidenberger, H.: How good are the visual MPEG-7 features?, Proc. of Visual Communication and Image Processing (VCIP), Lugano, Switzerland, (2003)
3. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7, Wiley (2002)
4. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Academic Press, (2003)
5. Won, C.S., Park, D.K., Park, S.-J.: Efficient use of MPEG-7 edge histogram descriptor, ETRI Journal, vol. 24, No. 1, (2002) 23–30

Automatic Synthesis of Background Music Track Data by Analysis of Video Contents

Toshio Modegi

Research & Development Center, Dai Nippon Printing Co., Ltd.
250-1, Wakashiba, Kashiwa-shi, Chiba 277-0871, Japan
Modegi-T@mail.dnp.co.jp
http://www.dnp.co.jp

Abstract. This paper describes an automatic creation technique of background music track data for given video file. Our proposed system is based on a novel BGM synthesizer, called "Matrix Music Player", which can produce 3125 kinds of high-quality BGM contents by dynamically mixing 5 audio files, which are freely selected from total 25 audio waveform files. In order to retrieve an appropriate BGM mixing patterns, we have constructed an acoustic analysis database, which records acoustic features of total 3125 synthesized patterns. Developing a video analyzer which generates image parameters of given video data and converts them to acoustic parameters, we access the acoustic analysis database and retrieve an appropriate synthesized BGM signal, which can be included in the audio track of the source video file.

1 Introduction

In general video productions, BGM or background music clips are often inserted as an acoustic effect with vocal tracks such as talk and narration data. In case of drama or documentary video production, the included music contents are artistically selected or newly composed by acoustic designers, according to story progression of the entire video program. However, in cases of background video, education program, information program and computer graphics animation oriented program productions, not so much attention must be paid for background music selections rather so much attention need to paid for production cost effectiveness. The target of this work is focusing on the latter cases of video productions, and reduces total production costs by developing an automatic audio editing system. This can select and insert automatically suitable BGM contents to producing video contents in as low copyright costs as possible.

There have been several works of adding automatically appropriate video clips to music contents [1], or analyzing and visualizing acoustic signals using computer graphics. But the other research approach of adding appropriate audio contents to video contents like what we are proposing has not been almost reported. This work is based on our developed mass production system of BGM contents called "Matrix Music Player"[2] as shown in Fig. 1. In this paper,

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 591–598, 2004.

additionally we propose an intelligent BGM retrieval technology of finding appropriate BGM patterns by acoustic feature parameters, and a video analysis technology of analyzing image feature parameters and converting them to acoustic parameters for BGM retrieval.

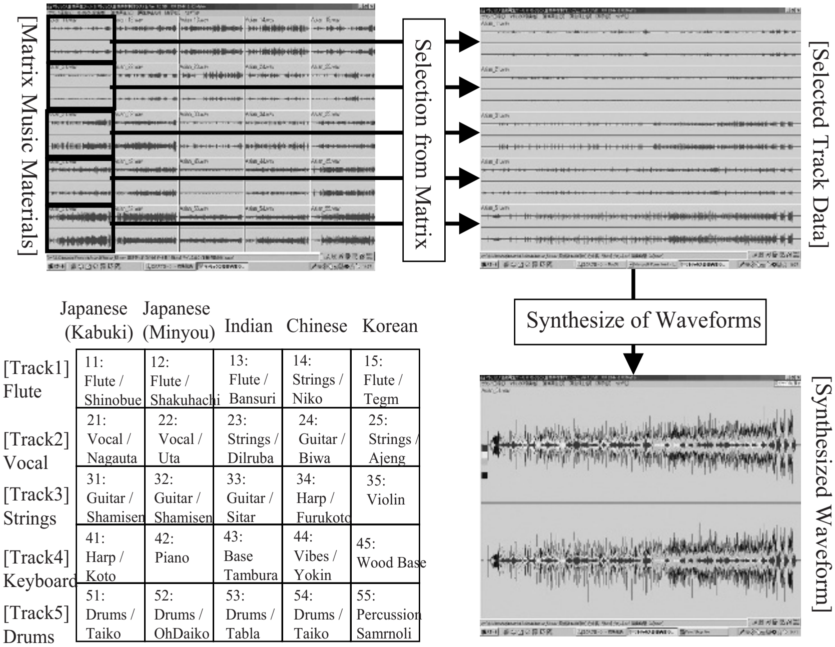
2 Music Synthesizing Technology of BGM Mass Production

Music works are composed of at least three parts: melodies, chords and rhythms. In general music compositions, each part is further divided to multiple instrument parts, whose editing tracks are used for source recording. If we can produce M sets of N-track music work whose some track data can be freely exchanged to the other track without making any musical contradiction; we can produce M^N number of different music contents by mixing N-track data randomly selected from M x N basic track data files. This music mass production concept, what we are proposing and calling as "Matrix Music Player", is especially useful for BGM content generation.

If we prepare 5 sets of 5-track music works, total 25 matrix format music data, like shown in Fig. 1, we can produce 3125 number of different music contents. In case of Fig. 1, each track corresponds with an instrument category: Flute, Vocals, Strings, Keyboard and Drums, and each track has 5 different voices of instrument used in 5 Asian cultures: Japanese (for dance), Japanese (for folksong), Indian, Chinese and Korean. Different from MIDI music, these 5 voices on the same track are not always based on the same music notes, but they are produced independently and can include vocal sounds. These track voices have the same length of playback time and are digitized as a high-quality stereo audio waveform format. We can produce even high-definition audio BGM contents by mixing 5-track waveform data from the beginning. In our prototype system of "Matrix Music Player", each audio track waveform has been sampled by 96kHz/24bits/2-ch and had 3-minute playback time. Moreover, if we compress each unit waveform file by lossless, we can archive total 25 CD-quality 6-minute waveform files into a single CD-ROM.

3 Intelligent Retrieval of BGM Synthesizing Patterns

In order to find appropriate BGM synthesizing patterns or matrix patterns from total 3125 possible patterns, we have provided an acoustic matrix at the front-end of the material matrix. In this acoustic matrix, we can define 8 parameters of acoustic features: volume, stereo, pitch, note, harmony, overtone, tempo and rhythm. In Fig. 2 example, we can define to each parameter four-level weight: high, middle, low and not considered, for each acoustic parameter. Before executing retrieval processes, we need to prepare an acoustic analysis database, in which 3125 records of acoustic parameter analysis data are archived. This database is created by analyzing 3125 patterns of BGM synthesized waveforms



	Japanese (Kabuki)	Japanese (Minyou)	Indian	Chinese	Korean
[Track1] Flute	11: Flute / Shinobue	12: Flute / Shakuhachi	13: Flute / Bansuri	14: Strings / Niko	15: Flute / Tegm
[Track2] Vocal	21: Vocal / Nagauta	22: Vocal / Uta	23: Strings / Dilruba	24: Guitar / Biwa	25: Strings / Ajeng
[Track3] Strings	31: Guitar / Shamisen	32: Guitar / Shamisen	33: Guitar / Sitar	34: Harp / Furukoto	35: Violin
[Track4] Keyboard	41: Harp / Koto	42: Piano	43: Base Tambura	44: Vibes / Yokin	45: Wood Base
[Track5] Drums	51: Drums / Taiko	52: Drums / OhDaiko	53: Drums / Tabla	54: Drums / Taiko	55: Percussion Samrnoli

Fig. 1. BGM Mass Production System Using "Matrix Music Player"

using the following formula and storing 8 kinds of calculated parameters for each pattern. Using this database we can search several appropriate BGM synthesized patterns corresponding with our specified acoustic matrix. Moreover, creating a Transform Knowledge Database, which transforms a certain keyword to a set of acoustic matrix parameters, we can search BGM patterns by defining sensitive or *kansei* keywords such as "cheerful".

3.1 Calculations of Acoustic Parameters

The following two parameters are calculated to given waveform data: $X(i), i = 0, S-1$.

(1) Volume parameter (indicating a dynamic range of given acoustic signal)

$$P_v = 20 \bullet \log\{\sum_{i=0}^{S-1} |X(i)|\}/S. \tag{1}$$

(2) Stereo parameter (indicating a spatial range from the left to right side)

$$P_s = 20 \bullet \log\{\sum_{i=0}^{S/2-1} |R(i)|\} \bullet 2/S. \tag{2}$$

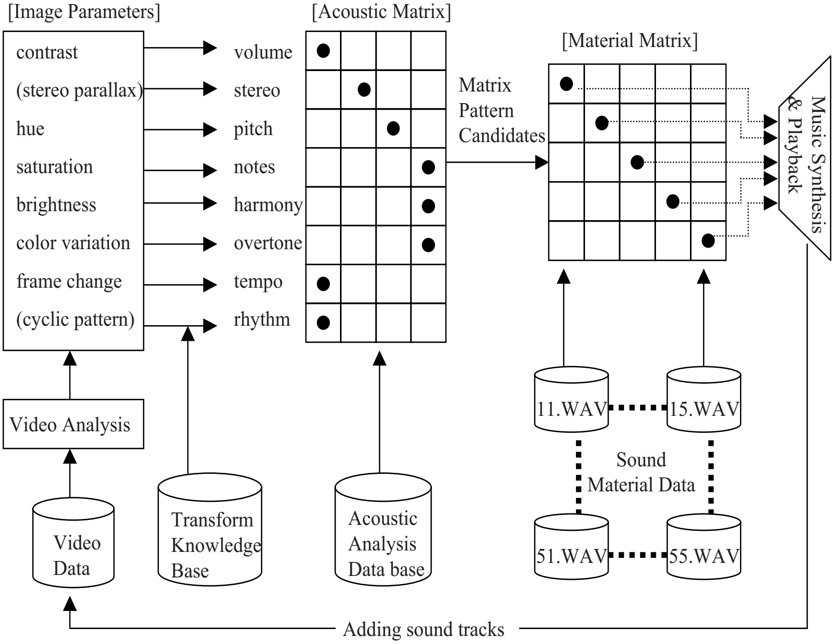


Fig. 2. BGM Sound Track Synthesis System for Video File

Here, R(i) is defined as follows:

```
If  |X(i*2)|>=|X(i*2+1)|  then  R(i)=X(i*2)/X(i*2+1).
If  |X(i*2)|<|X(i*2+1)|  then  R(i)=X(i*2+1)/X(i*2).
```

Next, calculating spectrogram: Zk(n), n=0,· · ·,N-1 (127), k=0,. . .,K-1, analyzed data to the waveform X(i), where n is MIDI note number indicating discrete frequency parameters and k is analyzed unit frame number; a frame size is 8192 samples. Using this spectrogram Zk(n), the following four parameters are obtained.

(3) Pitch parameter (indicating an average pitch of recorded sounds)

$$P_p = \left\{ \sum_{k=0}^{K-1} \frac{\sum_{n=0}^{N-1} nZ_k(n)}{\sum_{n=0}^{N-1} Z_k(n)} \right\} /K. \tag{3}$$

(4) Note parameter (indicating duplicated notes or number of ensemble instruments)

$$P_n = \left\{ \sum_{k=0}^{K-1} C(k) \right\} /K. \tag{4}$$

C(k) is a count of notes where $Zk(n)$ >*predetermined level* at a frame k.

(5) Harmony parameter (indicating whether the key is major or minor) Using m where $Zk(m)$ will be the maximum level at a frame, we define $P_h(k)$ as

$P_h(k) = \{Z_k(m+4) - Z_k(m+3) + Z_k(m+16) - Z_k(m+15) + Z_k(m-8) - Z_k(m-9)\}/6$. This calculation means addition of major third interval level (+4 semitone) and subtraction of minor third interval level (+3 semitone) including upper or lower octave interval level.

$$P_h = \{\sum_{k=0}^{K-1} P_h(k)\}/K. \tag{5}$$

(6) Overtone parameter (indicating richness of overtone)

$$P_o = \{\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} \frac{Z_k(n) + Z_k(n+12) + Z_k(n+19) + Z_k(n+24)}{4}\}/K. \tag{6}$$

Furthermore, temporally shrinking the waveform at 1/60 ratio, and calculating the spectrogram again to the shrunken waveform: Zk(n), n=0,· · ·N-1 (127), k=0,· · ·,L-1, we obtain the top two spectrum level of note number: M1 and M2 ($M1 < M2$) at each frame where analysis frame size is 8129. The following two parameters are obtained by these M1 and M2 value.

(7) Tempo parameter (indicating an average beat per minute, unit:BPM)

$$P_t = \{\sum_{k=0}^{L-1} 440 \bullet 2^{\frac{M2-64)}{12}}\}/L. \tag{7}$$

(8) Rhythm parameter (indicating an average rhythm count)

$$P_r = \{\sum_{k=0}^{L-1} 100 \bullet 2^{\frac{M1-M2}{12}}\}/L. \tag{8}$$

3.2 Calculations of Spectrogram

A spectrogram: Zk(n), n=0,· · ·,N-1 (127), k=0,· · ·,K-1, analyzed data to the waveform $X(i)$ is calculated by the following, where n is MIDI note number indicating discrete frequency parameters and k is analyzed unit frame number.

$$Z_k(n) = [\{\sum_{i=kT}^{kT+T-1} Xa(i)C_n(i)\}^2 + \{\sum_{i=kT}^{kT+T-1} Xa(i)S_n(i)\}^2]^{1/4}. \tag{9}$$

Here, $C_n(i) = \cos(2\pi f(n)i/F)$, $S_n(i) = \sin(2\pi f(n)i/F)$, F is a source sampled frequency value, f(n) means MIDI note number frequency, $Xa(i) = \{X(i*2) + X(i*2+1)\}/2$ and $f(n) = 440 \bullet 2^{(n-69)/12}$. For calculating both tempo and rhythm parameters, we calculate a spectrogram to the 1/60 temporally shrunken waveform X(i). In this case, the unit of frequency becomes BPM from Hz.

$$Z_k(n) = [\{\sum_{j=kT}^{kT+T-1} Y(j)C_n(j)\}^2 + \{\sum_{j=kT}^{kT+T-1} Y(j)S_n(j)\}^2]^{1/4}. \tag{10}$$

Here $Y(j) = \sum_{i=0}^{119} X(j*120+i)$.

4 Analysis of Source Video Images

By analyzing each image frame in given video file, we calculate 8 image parameters: contrast, stereo parallax, hue, saturation, brightness, color variation, frame change, cyclic pattern. Then each image parameter will be converted to the specified acoustic parameter described before, based on the transform knowledge database as shown in Fig. 2. As described before, by executing an intelligent retrieval using these converted acoustic parameters, we can obtain several appropriate BGM synthesized patterns. Synthesized BGM audio data can be inserted in the sound tracks of the analyzed video file. The two image parameters of stereo parallax and cyclic pattern are applied only for 3-D video images and loop based video images, therefore these two parameters are not used for general video files.

4.1 Calculations of Image Parameters

Defining RGB pixel value at (x,y) position on a frame f in given video file as R(f,x,y), G(f,x,y) and B(f,x,y), each pixel values are converted to the HSV color space as H(f,x,y), S(f,x,y) and V(f,x,y). Using these 6 pixel values, 6 image parameters excluding stereo parallax and cyclic pattern parameters are calculated and they are converted 6 three-level acoustic parameters as follows.

(1) Color variation parameter

Compressing gradation level of R(f,x,y), G(f,x,y) and B(f,x,y) to 16 levels, A number of RGB color variation included in frame f is counted, then an average color variation value of total frames is converted to an overtone acoustic parameter. The maximum value of color variation image parameter is 4096, and this value is converted to a three-level acoustic overtone parameter based on the transform knowledge database.

(2) Frame change parameter

Calculating all of pixel value distance between R(f,x,y), G(f,x,y) and B(f,x,y) in a frame f, and R(f-1,x,y), G(f-1,x,y) and B(f-1,x,y) in a previous frame $f-1$ as follows.

$$D(x,y) = \sqrt{D_r(x,y)^2 + D_g(x,y)^2 + D_b(x,y)^2}. \tag{11}$$

Here $D_r(x,y) = R(f,x,y) - R(f-1,x,y)$, $D_g(x,y) = G(f,x,y) - G(f-1,x,y)$ and $D_b(x,y) = B(f,x,y) - B(f-1,x,y)$. An average distance value of total pixels in total frames is converted to a three-level acoustic tempo parameter.

(3) Hue, Saturation and Brightness parameters

Average pixel values of H(f,x,y), S(f,x,y) and V(f,x,y) of total pixels in total frames are calculated. The average hue value, the average saturation value and the average brightness value are converted to three-level acoustic pitch, notes and harmony parameters, respectively.

(4) Contrast parameter

Finding both the maximum pixel value V_{max} and the minimum pixel value V_{min} of brightness V(f,x,y) in a frame f , and a contrast value C is defined as $C = V_{max} - V_{min}$. An average contrast value C of total frames is converted to a three-level acoustic volume parameter.

4.2 Calculations of Multiple Sets of Image Parameters

General video files are composed of a lot of different scenes, and providing a single type of BGM audio stream to a single video file is not always appropriate. Our proposing image analysis described before is executed on frame by frame basis, and a set of image parameters for each frame is calculated. In the previous section, we calculated average image parameter values of whole frames, moreover we can divide a video file to several groups of frames and calculate several sets of image parameters corresponding with divided groups. For dividing files, we define slice levels for image parameter, and divide files where image parameter difference values between nearest two frames are larger than the predefined slice levels. If the image parameters in the frame K is prominently different from the image parameters in the frame K+1, the video file is divided to two parts before and after the frame K.

5 Experimental Results

We have developed a real-time BGM synthesizing software running on Microsoft Windows PC based on Fig. 2, and have tried BGM synthesis experiments using several around 20-second video clips as shown in Fig. 3. The used video files have been selected from commercially available loyalty free video contents, "Pick-a-Video Pro, DV format series," produced by Ulead Systems. Each video clip has been analyzed by the method described before at almost real time and 6 image

	Video frame				
Image Parameters	contrast	168	132	260	11
	hue	85	88	25	118
	saturation	161	147	175	48
	brightness	252	236	229	254
	color variation	149	150	105	150
	frame change	56	56	20	34
Acoustic Parameters	volume	3	3	3	3
	pitch	2	2	3	1
	notes	2	2	1	2
	harmony	3	2	3	1
	overtone	2	2	2	2
	tempo	2	2	1	1
BGM Matrix	track1	□□●□□	□□□●□	□□●□□	□□●□□
	track2	□□●□□	□□●□□	□●□□□	●□□□□
	track3	●□□□□	□□□□●	□●□□□	□□●□□
	track4	●□□□□	●□□□□	□□□●□	●□□□□
	track5	□□□●□	□□●□□	□□●□□	□●□□□

Fig. 3. Results of BGM Synthesis Experiments Using 20-second Video Clips

parameters, whose range has from 0 to 255, have been calculated, and each image parameter has been converted to three-step value of its corresponding acoustic parameter. Then using sets of converted acoustic parameters, intelligent retrieval processes have been executed, and several BGM matrix patterns have been extracted, and in Fig. 3 only the top patterns of extracted candidate matrix patterns for each video clip are shown. A set of BGM matrix pattern is composed of 5-level of 5 values indicating waveform file selections for 5 tracks, as described in Fig. 1.

6 Conclusions

We have proposed and developed a BGM automatic synthesizing software for video clips, and have tried BGM synthesis experiments using several around 20-second video clips. The given video clips could be analyzed, and some BGM audio streams could be automatically synthesized and be inserted to the given video files. The inserted BGM audio streams for all of the given video clips have been objectively acceptable.

In the future we will support multiple sets of image parameter analysis for a single video file. We will develop a real-time audio content switching technique for streaming video contents that scene changes in video stream can be detected at real time and appropriate BGM audio streams can be smoothly selected and switched.

This research work has been planned with Matrix Music Corporation, and promoted by the Digital Content Association of Japan as the 2003-year government project "Development of Break Through Technologies on Broadband Contents", being also financially supported by the Economy, Trade and Industry Japan. The author gives thanks for all of the persons concerned in this project including the persons in Communication & Information Operation in our company and DNP Information Systems Co., Ltd.

References

1. Foote, J., Cooper, M., Girgensohn A.: Creating music videos using automatic media analysis. Proceedings of the tenth ACM international conference on Multimedia, New York (2002) 553–560
2. Modegi, T.: Design of Network Based Multi-track Audio Player System. Proceedings of IEICE General Conference,A-10-3, Tokyo (2004) 233–233
3. Modegi, T.: Development of Lossless Encoder Tool and Production of a High Definition Music Work for Evaluation. The 50-th meeting of IPSJ Musical Informatics SIG, MUS-50-2, Tokyo (2003) 7–12

Picture Quality Improvement in Low Bit Rate Video Coding Using Block Boundary Classification and Simple Adaptive Filter

Kee-Koo Kwon, Jin-Suk Ma, Sung-Ho Im, and Dong-Sun Lim

Embedded S/W Technology Center, ETRI
161 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Korea
{kwonkk,jsma,shim,dslim}@etri.re.kr

Abstract. In this paper, we propose a novel post-filtering algorithm with low computational complexity that improves the visual quality of block-based coded images using block boundary classification and simple adaptive filter (SAF). In this algorithm, each block boundary is classified into smooth or complex sub-region, and the existence of blocking artifacts is determined using blocky strength for smooth-smooth sub-regions. Simple adaptive filtering is processed adaptively in each block boundary. That is, a nonlinear 1-D 8-tap filter is applied to smooth-smooth sub-regions with blocking artifacts, and for smooth-complex or complex-smooth sub-regions, a nonlinear 1-D variant filter is applied to block boundary pixels so as to reduce the blocking and ringing artifacts. And for complex-complex sub-regions, a nonlinear 1-D 2-tap filter is only applied to adjust two block boundary pixels so as to preserve the image details. Experimental results show that the proposed algorithm produced better results than those of the conventional algorithms both subjective and objective viewpoints.

1 Introduction

Block DCT-based coding techniques have been adopted in many international standards, including H.263+ [1] and MPEG-4 [2], [3]. However, such techniques produce noticeable blocking artifacts along block boundaries and ringing artifacts or mosquito noise near edges in decompressed images at low bit rates, because the DCT coefficients in each block of an image are processed and quantized independently [1]-[6]. Therefore an efficient deblocking and deringing scheme is essential for preserving the visual quality of decompressed images.

A variety of post-filtering schemes for reducing the blocking and ringing artifacts have already been proposed to improve the visual quality of block-based coded images in the decoder, such as adaptive filtering methods in the spatial domain [1]–[3], the projections onto convex sets (POCS)-based method [4], estimating the lost DCT coefficients in the transform domain [5], and wavelet transform-based method [6]. Among these algorithms, the spatial domain filtering methods have the advantage of simplicity and easy hardware implementation

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 599–606, 2004.

[1]–[3]. In H.263+, the blocking artifacts are reduced using the loop filter and post filter [1]. In MPEG-4 committee draft (CD), the blocking and ringing artifacts are reduced using the verification model (VM) post filter, deblocking filter and deringing filter [2], [3]. The MPEG-4 deblocking algorithm which operates in two modes: DC offset mode for low activity blocks and default mode. Block activity is determined according to the amount of changes in the pixels near the block boundaries. All modes apply a 1-D filter in a separable way. The default mode filter uses the DCT coefficients of the pixels being processed and the DC offset mode uses a 1-D filter. Although this algorithm can conserve the complex regions, it is unable to eliminate the blocking artifacts in complex regions efficiently.

Accordingly, this paper proposes a novel post-filtering algorithm to reduce the blocking and ringing artifacts in low bit rate block-based coded images. The major objective of our algorithm is to develop a post-filtering technique that is applicable in a real-time decoding system. We attempt to find a good algorithm that has a low computational complexity and good visual quality of decoded images.

In this algorithm, each block boundary is classified into smooth or complex sub-region using the statistical characteristics of four pixels within a block boundary and the existence of blocking artifacts in smooth-smooth sub-regions is determined using blocky strength, the difference between two pixels within a block boundary. And simple adaptive filtering is processed in each block boundary adaptively, that is, a nonlinear 1-D 8-tap filter is applied to smooth-smooth sub-regions, and for smooth-complex or complex-smooth sub-regions, a nonlinear 1-D variant filter is applied to four block boundary pixels in smooth sub-region and block boundary pixel only in complex sub-region so as to reduce the blocking and ringing artifacts simultaneously. And for complex-complex sub-regions, a nonlinear 1-D 2-tap filter is only applied to adjust two block boundary pixels so as to reduce the blocking artifacts and preserve the image details.

Experimental results show that the proposed algorithm improved the PSNR and visual quality of MPEG-4 decoded sequences, and produced better results than those of the conventional algorithms.

2 Proposed Post-filtering Algorithm

2.1 Detection of Blocking Artifacts and Block Boundary Classification

In this algorithm, as shown in Fig. 1, each block boundary is classified into smooth or complex sub-region using the statistical characteristics of four pixels within a block boundary as follows:

$$\begin{array}{ll} \textit{if}\ (var_n < T_s) & \textit{smooth sub-region} \\ \textit{else} & \textit{complex sub-region} \end{array}$$

where, var_n is the local variances of four pixels in each block boundary, n=1, 2, $T_s = 0.5\times$QP, and QP is a quantization parameter. And each block boundary is classified into smooth-smooth sub-regions, smooth-complex or complex-smooth

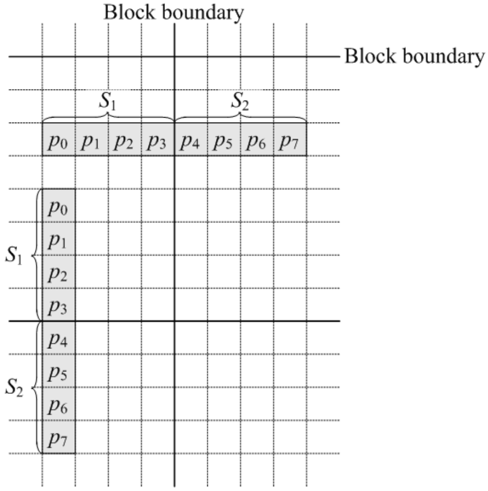


Fig. 1. The eight pixels used for block boundary classification and smooth-smooth sub-region filtering

sub-regions, and complex-complex sub-regions. In case of smooth-smooth sub-regions, there exist the regions those are not existed the blocking artifacts, so the existence of blocking artifacts is detected and then adaptive filtering is processed to reduce the computational complexity. So we use the blocky strength B_s, the difference between two pixels within a block boundary, to detect the blocking artifacts, that is,

$$\begin{array}{ll} \textit{if } (|B_s| \geq T_m) & \textit{blocking artifacts are existed} \\ \textit{else} & \textit{blocking artifacts are not existed} \end{array}$$

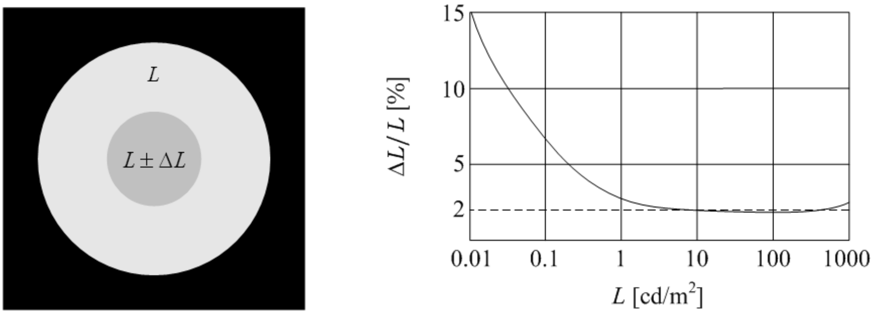
where, $B_s = p_3$ - p_4 and $T_m = 3$. The threshold T_m is determined using Weber's law [7] as shown in Fig. 2. Weber's law states that the ratio of the increment threshold to the background intensity is a constant in $10 \sim 10^3$ cd/m^2 intensity range. That is,

$$\Delta L/L \approx 0.02 \; . \tag{1}$$

So the threshold T_m is determined considering the intensity variation of mid-range intensity 127 in 8 bits image (that is, 127×0.02≈3) and filtering complexity.

2.2 Adaptive Block Boundary Filtering

Based on the above classification scheme, a novel post-filtering method using spatial adaptive filter is proposed. Because human visual system (HVS) is sensitive to the blocking artifacts in smooth region more than complex region, the proposed method processed adaptively using three different filter, a nonlinear 1-D 8-tap filter, a nonlinear 1-D variant filter, and a nonlinear 1-D 2-tap filter.



(a) (b)

Fig. 2. (a) Experimental environment and (b) the resulted characteristic curve of the Weber's law

Table 1. The relationship between quantization parameter QP and T_c

QP	1~6	7~9	10~12	13~15	16~18	19~21	22~24	25~27	28~30	31
T_c	1	2	3	4	5	6	7	8	9	10

For smooth-smooth sub-regions with the blocking artifacts, 1-D 8-tap spatial adaptive filter is used to process the block boundary as shown in Fig. 1. The outputs of the deblocking filter are formed as

$$\widehat{p}_n = p_n + sign(B_s)\frac{B_s}{\alpha_n}, n = 0, 1, \cdots, 7 \tag{2}$$

where, $\alpha_n = \{8, 6, 3, 2, -2, -3, -6, -8\}$. In smooth-smooth sub-regions, the blocking artifacts seem a step-wise function, so the blocking artifacts can be reduced by above simple scheme.

And for smooth-complex or complex-smooth sub-regions, a nonlinear 1-D variant filter is applied to four block boundary pixels in smooth sub-region and block boundary pixel only in complex sub-region so as to reduce the blocking and ringing artifacts simultaneously. The proposed method is filtered adaptively by using blocky strength B_s and QP. The relationship QP and threshold value T_c is given in Table 1. And the outputs of the filter in smooth-complex sub-regions are formed as

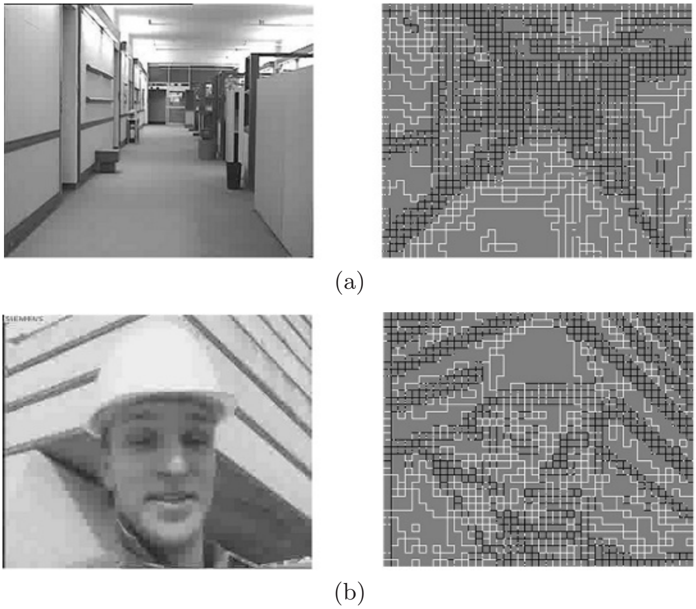
i) $|B_s| \leq 2 \times QP$

$$\widehat{p}_3 = p_3 + sign(B_s)\frac{B_s}{K} \tag{3}$$

$$\widehat{p}_4 = p_4 - sign(B_s)\frac{B_s}{K} \tag{4}$$

ii) $|Bs| > 2 \times QP$

no filtering



(a)

(b)

Fig. 3. Sequences with CIF decoded by MPEG-4 and results of block boundary classification for each sequence: (a) Hall monitor (QP=12); (b) Foreman (QP=30)

where, $K = 4$. But if above scheme is applied to all block boundary pixels, the blocking artifacts are occurred in the inner parts of the block and the image is over-smoothed. So we process the following step.

$$\Delta p_n = \frac{1}{2}\left[clip(p_n - p_{n-1}) + clip(p_n - p_{n+1})\right], n = 3, 4 \tag{5}$$

$$\hat{p}_n = p_n - \Delta p_n, n = 3, 4 \ . \tag{6}$$

Here, $clip(x)$ is as follows:

$$clip(x) = \begin{cases} -T_c, & x < -T_c \\ x \ , & |x| \leq T_c \\ T_c, & x > T_c \ . \end{cases} \tag{7}$$

The $clip(\cdot)$ process is used to prevent the over-smoothing. And the pixels p_0, p_1, and p_2 in smooth sub-region are processed in the same way as the above scheme. And for complex-smooth sub-regions, the pixels p_3, p_4, p_5, p_6, and p_7 are processed in the same way as the above scheme.

And for complex-complex sub-regions, a nonlinear 1-D 2-tap filter is only applied to adjust two block boundary pixels p_3 and p_4 so as to reduce the blocking artifacts and preserve the image details. The processing method are similar to above method, but the value $K = 8$. After the horizontal block boundary is processed, the vertical block boundary is processed in the same way as the horizontal block boundary.

Table 2. Experimental results for MPEG-4 decoded sequences

Bit rates, size, frame rates	Sequences	QP	Average PSNR [dB]			
			MPEG-4 decoded	VM-18 post filter		Proposed post filter
				Deblocking only	Both filtering	
10 kbps, QCIF, 7.5 Hz	Container ship	17	29.53	29.78	29.71	29.74
	Moth. & daugh.	15	32.28	32.42	32.44	32.47
24 kbps, QCIF, 10 Hz	Container ship	10	32.78	32.96	32.89	32.94
	Moth. & daugh.	8	35.32	35.38	35.28	35.46
48 kbps, QCIF, 10 Hz	Foreman	13	30.96	31.10	31.04	31.10
	Coastguard	14	29.08	29.11	28.96	29.08
	Silence voice	7	34.42	34.59	34.63	34.62
48 kbps, CIF, 7.5 Hz	News	18	31.26	31.45	31.37	31.49
	Moth. & daugh.	10	36.10	36.12	36.01	36.23
	Hall monitor	12	33.60	33.82	33.60	33.97
112 kbps, CIF, 15 Hz	News	11	34.16	34.35	34.27	34.40
	Foreman	30	28.43	28.55	28.53	28.53
	Coastguard	29	26.53	26.55	26.37	26.46

3 Experimental Results

To evaluate the performance of the proposed algorithm, computer simulations were performed using MPEG-4 VM 18.0 low bit rate video coder [2]. Each video sequence is QCIF (176×144) or CIF (352×288) in size, 300 frames, and compressed at various bit rates.

The proposed block boundary classification method classifies efficiently for various bit rates decoded images. Figure 3 shows that the results of the proposed block boundary classification method.

To compare the proposed algorithm with MPEG-4 VM-18 post filter [2], the PSNR performances are presented in Table 2 and Fig. 4. In Table 2, the proposed algorithm produced the maximum 0.37 PSNR improvement than those of VM-18 post filter. And in Fig. 4, the proposed algorithm produces better result than the performances of the VM-18 post filter for all frames. Coastguard sequence decoded by MPEG-4 with 112 kbps, 15 Hz, QP=29, and the post-processed sequences are shown in Fig. 5. The proposed algorithm effectively reduced the blocking artifacts and preserved the original high-frequency components, such as edges. But, VM-18 post filter reduced the blocking artifacts in smooth regions, but the blocking artifacts in complex regions are still remained.

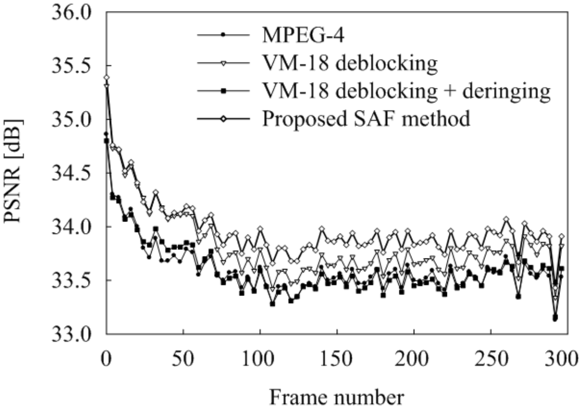


Fig. 4. Experimental results by VM-18 post filter and proposed post filter method for Hall monitor sequence decoded by MPEG-4 with CIF, 48 kbps, 7.5 Hz frame rate, QP=12

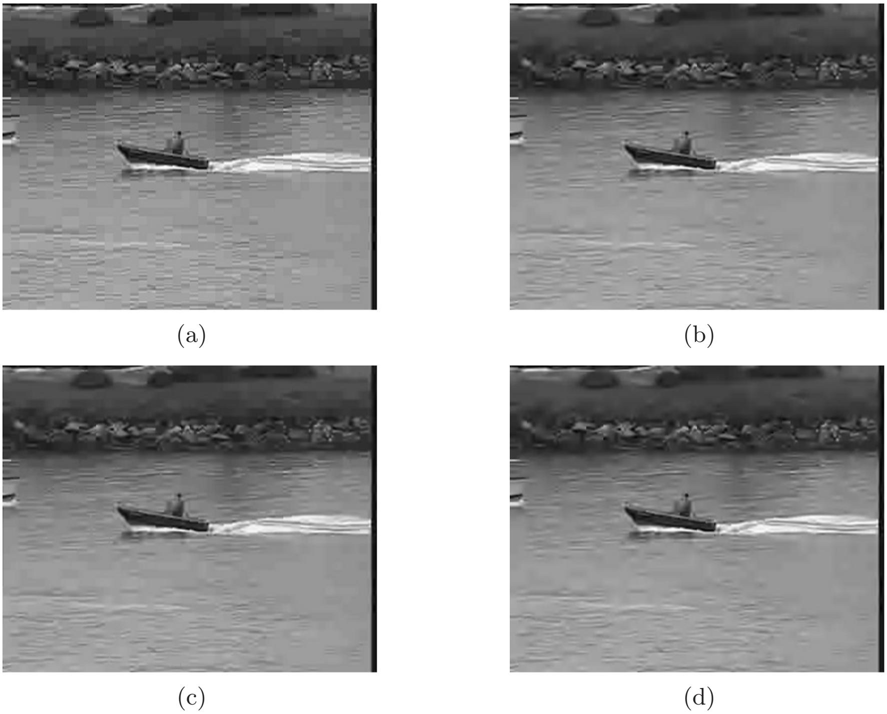


Fig. 5. (a) The Coastguard sequence decoded by MPEG-4 with 112 kbps, 15 Hz, QP=29, and post-processed sequences by (b) VM-18 deblocking filter method, (c) VM-18 deblocking and deringing filter method, and (d) proposed SAF method

4 Conclusions

We propose a novel post-filtering algorithm with low computational complexity that improves the visual quality of decoded images using block boundary classification and simple adaptive filter. In this algorithm, each block boundary is classified into smooth or complex sub-region and the existence of blocking artifacts in smooth-smooth sub-regions is determined. And simple adaptive filtering is processed in each block boundary adaptively, that is, a nonlinear 1-D 8-tap filter is applied to smooth-smooth sub-regions, and for smooth-complex or complex-smooth sub-regions, a nonlinear 1-D variant filter is applied to four block boundary pixels in smooth sub-region and block boundary pixel only in complex sub-region. And for complex-complex sub-regions, a nonlinear 1-D 2-tap filter is only applied to adjust two block boundary pixels so as to reduce the blocking artifacts and preserve the image details. Experimental results show that the proposed algorithm produced the maximum 0.37 PSNR improvement than those of conventional algorithms and subjective quality is better than those of the conventional algorithms, especially the complex regions.

References

1. ITU-T Recommendation H. 263 Version 2: Video Coding for Low Bit Rate Communication. Draft (1998)
2. MPEG Video Group: MPEG-4 video verification model version 18.0. ISO/IEC JTC1/SC29/WG11 N3908 (2001)
3. Kim, S. D., Yi, J. Y., Kim, H. M., and Ra, J. B.: A deblocking filter with two separate modes in block-based video coding. IEEE Trans. Circuits Systems Video Technol. 9 (1999) 156–160
4. Yang, Y., Galatsanos, N., and Katsagelos, A.: Projection-based spatially adaptive reconstruction of block-transform compressed images. IEEE Trans. Image Processing. 4 (1995) 896–908
5. Paek, H., Kim, R. C., and Lee, S. U.: A DCT-based spatially adaptive post-processing technique to reduce the blocking artifacts in transform coded images. IEEE Trans. Circuits Syst. Video Technol. 10 (2000) 36–41
6. Xiong, Z., Orchard, M. T., and Zhang, Y. Q.: A deblocking algorithm for JPEG compressed images using overcomplete wavelet representations. IEEE Trans. Circuits Syst. Video Technol. 7 (1997) 433–437
7. Jain, A. K.: Fundamentals of Digital Image Processing. Prentice Hall, New York (1990)

Bit Position Quantization in Scalable Video Coding for Representing Detail of Image

Hideaki Kimata, Masaki Kitahara, Kazuto Kamikura, and Yoshiyuki Yashima

NTT Cyber Space Laboratories, NTT Corporation,
1-1, Hikari-no-oka, Yokosuka, 239-0847, Japan
kimata@m.ieice.org

Abstract. We have studied a coding method to provide a different type of distortion from in the images of the base layer using a scalable coding scheme. Conventionally for the enhancement layer in the scalable video coding scheme, quantization in transformed domain on the residual coding reduces much high frequency information of images. The decoded images are basically blurred, i.e. it lacks detail information. This paper proposes a novel residual coding method to add detail information upon the decoded images of the base layer. The proposed method uses the BPQ, which is one of non-linear scalar quantization methods and which represents more efficiently the feature of detail information than a quantization method in DCT domain. And the proposed method uses the appropriate entropy coding method for the q uantized values. The proposed entropy coding uses the lossless coding method extended f rom the H.264/AVC.

1 Introduction

Scalable video coding is highly expected for video communication or distribution service over various transmission QoS networks such as the Internet. Basically, coding efficiency of the scalable coding method is worse than that of rate-distortion optimized single layer coding method. Because of this fact, we have studied a scalable video coding from another viewpoint. Objective of our study is to provide a different type of distortion from in the images of the base layer using a scalable coding scheme. In this paper, we propose a coding method of the enhancement layer, which does not use DCT, for the purpose of efficiently coding high frequency of the images.

MPEG-4 Simple Scalable Tool and MPEG-4 FGS are SNR scalable coding methods of the enhancement layer. Especially FGS provides fine grain SNR scalability. These use DCT to encode the residual data of the enhancement layer. In the coding scheme using DCT, high frequency coefficients in DCT domain are usually cut off by non-linear quantization for the purpose of improving coding efficiency. This method successfully decreases the average of differences of the original images and the decoded images, but the decoded images are blurred, because high frequency of images that is detail information of images is reduced. The same problem exists for the base layer when the quantization in transform

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 607–614, 2004.

domain is used. For instance, H.264/AVC uses DCT to encode the residual data in prediction coding. In addition, to reduce the quantization noise, the deblocking filter is applied. This filtering reduces the high frequency information more, so the decoded images are more blurred.

Quantization in pixel domain is one solution to encode high frequency information efficiently more than quantization in DCT domain. A scalar quantization can improve detail information at pixel level. In the simple scalar quantization, because the value is just scaled by the preset parameter, the quantized values are not highly correlated among adjacent pixels, therefore coding efficiency of entropy coding of the quantized values is low. Another approach is generating missing high frequency information at the decoder. In [1], the decoder receives the edge position information coded with MPEG-4 binary shape coding, and it generates high frequency information at the edge positions by enhancing the contrast. However, these methods are not enough to generate high frequency information because additionally transmitted data is only binary information.

In this paper, we propose a residual coding method of the enhancement layer. The proposed method uses a novel non-linear quantization method in pixel domain, bit position quantization (BPQ). The BPQ is developed for representing detail information more efficiently than the quantization in DCT domain in the same distortion (PSNR) level. Differently from the normal scalar quantization in pixel domain, the BPQ outputs the data that express the significant bit position of residual data in the range of predetermined positions. Because the significant bit represents the magnitude of the residual data well, the BPQ improves coding efficiency more than the normal scalar quantization. We also propose the efficient entropy coding methods of the BPQ data.

2 Bit Position Quantization

The BPQ treats residual data as an input data. The block diagram of coding residual data applying the BPQ is illustrated in Figure 1. The residual data is quantized based on the bit position of itself with the BPQ When 8 bit original image is coded in the base layer, the residual data is in the range of +/- 255, which is represented by the 8 bit for the absolute number and 1 bit for sign information. The most significant bit position of "1" in the absolute number with sign information is the output of the BPQ. And we preset the range of bit position to be coded, as the quantization parameter. Flow of the BPQ is as follows. The absolute number is clipped to 2^n, where "n" is equal or greater than zero. The bit position "bp" is measured from LSB, and then it is equal to (n+1). For instance, if the residual data is -36 in decimal, the binary of the absolute number is 100100. Then the BPQ outputs the value of 6 with negative sign information. In this paper, "n_min" denotes the minimum value of "bp" for negative sign and "p_min" denotes the minimum value of "bp" for positive sign, which are preset range of bit position as quantization parameter.

And the decoder side, inverse BPQ is applied. In the inverse BPQ, the median absolute number "msn" and sign information are obtained from the decoded bit

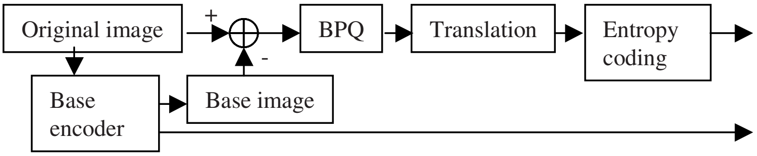


Fig. 1. Block diagram of coding of the residual data

position "m". The "msn" is the median value of the maximum and the minimum value with the same most significant bit position of "1". The "msn" is added to base images and decoded images are obtained. The value of "msn" is as follows.

$$msn = (2^{m-1} + 2^m - 1)/2 \tag{1}$$

3 Coding Method of BPQ Data

3.1 Translation of BPQ Data

The outputs of the BPQ are the bit position of the median absolute number and sign information. The value of "bp" and sign are translated into one symbol "ts" per a pixel. In the proposed translation method, "ts" is calculated by,

$$\begin{aligned}
&ts = 0 \qquad for\ bp < m1\ and\ s > 0\ (that\ is\ bp = 0)\\
&ts = bp - m1 + 1 \qquad for\ m1 <= bp < m2\ and\ s > 0\\
&ts = 2(bp - m2) + (m2 - m1) + E(s < 0) \qquad for\ others\,,
\end{aligned} \tag{2}$$

where "s" is calculated by the following.

$$s = sign(n\_min - p\_min) \tag{3}$$

When $n\_min > p\_min$, "$m1$" and "$m2$" are set to $m1 = p\_min$ and $m2 = n\_min$, otherwise set to $m1 = n\_min$ and $m2 = p\_min$. $E(s < 0)$ is set to 1 when $s < 0$, otherwise it is set to zero. Since LSB bits represent small difference from base images to original images, the value of "bp" corresponding to LSB bits should be translated into small number of symbol "ts". The proposed translation method is called LC that means LSB centric method, in this paper.

For comparison discussed in the later section 4, one compared method is explained here. "ts" is calculated as the following. This compared method is called MC that means MSB centric method, in this paper

$$\begin{aligned}
&ts = 0 \qquad for\ bp < n\_min\ (that\ is\ bp = 0)\\
&ts = 1 - n\_min + bp \qquad for\ bp >= n\_min\ and\ sign > 0\\
&ts = 17 - n\_min - bp \qquad for\ bp >= n\_min\ and\ sign < 0
\end{aligned} \tag{4}$$

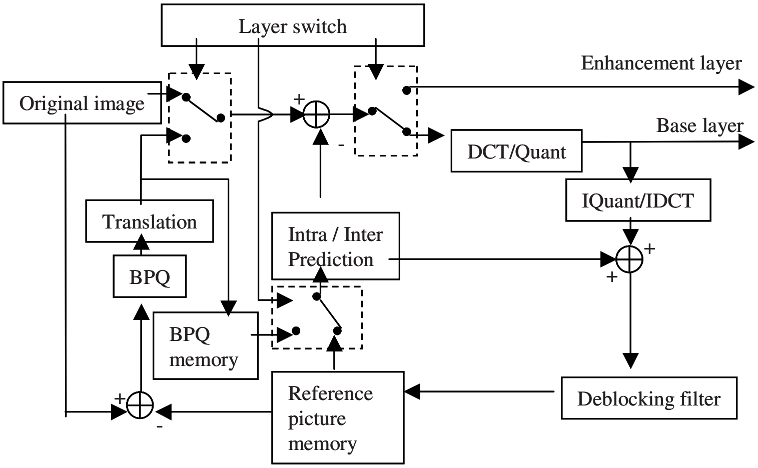


Fig. 2. AVC based encoder structure for the BPQ scalable coding

3.2 Entropy Coding of BPQ Data

After translation, entropy coding of "*ts*" is applied. Since "*ts*" is translated into per a pixel, two-dimensional array of "*ts*" is available for an image. Lossless coding method for two-dimensional images is possible to be applied for such data. We adopted the method that was proposed as extension of H.264/AVC [2]. In this paper we call this method AVC-LS. In AVC-LS, the encoder bypasses DCT and deblocking filter in the coding loop. AVC-LS has the same functionalities for predictive coding, such as intra prediction and reference picture selection methods. When the base layer is coded by H.264/AVC, both an encoder and a decoder are developed as an extension of H.264/AVC. Figure 2 illustrates the encoder structure of the extension of H.264/AVC. Basically two categories of reference picture memory are equipped, and one of memories is selected for the corresponding layer.

4 Evaluation

4.1 Comparison of BPQ and DCT

The effectiveness of the BPQ was evaluated in terms of PSNR and subjective quality compared with DCT based layered coding method. In this experiment, only luminance information is coded for the enhancement layer and compared as we set the assumption that chrominance information is sufficient in the base images. In the DCT based layered coding method, only luminance information of residual data is coded with P-picture coding type of H.264/AVC, where base images are applied as reference pic-tures. Table 1 lists the coding conditions of the base images that are coded with H.264/AVC and those of the BPQ that are coded

Table 1. Test conditions

Common conditions to H.264/AVC, and AVC-LS	Entropy coding	CABAC
	Framerate	30 fps
	Motion Compensation accuracy	1/4
	Mode decision	Optimization with Lagrangian
Only to H.264/AVC	QP	32, 28, 24
Test sequence	cheer_leader, coast, mobile, tempete (CIF, 150 frames)	

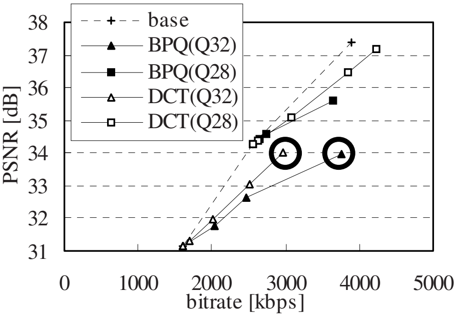
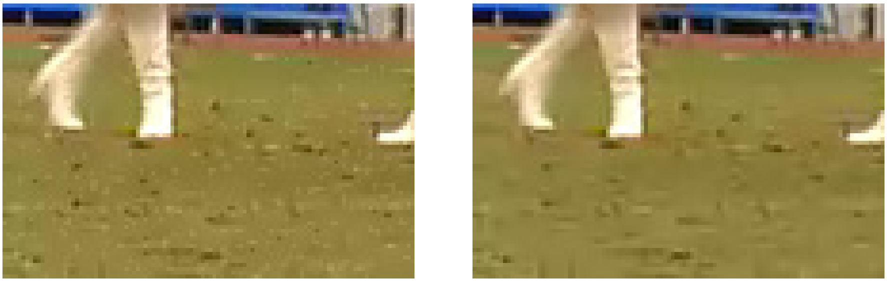


Fig. 3. PSNR of the BPQ and DCT methods

with AVC-LS, and it lists the test sequences. Three patterns of range of bit position "bp", i.e. quantization parameters for the BPQ were experimented, which are $R1 : bp = 8, 7, 6, 5(sign > 0), 8, 7, 6(sign < 0)$, $R2 : bp = 8, 7, 6, 5(sign > 0), 8, 7, 6, 5(sign < 0)$, $R3 : bp = 8, 7, 6, 5, 4(sign > 0), 8, 7, 6, 5(sign < 0)$. The bitrate of the residual data increases as the pattern of BPQ is changed from $R1$ to $R3$. The applied translation method is LC, and entropy coding of BPQ data is AVC-LS adaptive method. AVC-A denotes this AVC-LS adaptive method. AVC-I denotes all the pictures of residual data are coded with AVC-LS I-picture, and AVC-P1 denotes all the pictures except the first picture are coded with AVC-LS P-picture with a single reference picture, and AVC-P5 denotes all the pictures except the first picture are coded with AVC-LS P-picture with 5 reference pictures. In AVC-A, prediction type of picture is adaptively selected from I-picture, P-picture with a single reference picture, or P-picture with 5 reference pictures. Figure 3 shows the luminance PSNR at various bitrates for "*cheer_leader*". $BPQ(Qx)$ denotes the data with the BPQ method and $DCT(Qx)$ denotes the data with the DCT method, when QP of the base images is set to x. We can see that coding efficiency of the BPQ and the DCT are worse than the base layer with smaller QP. And we can also see that PSNR of the BPQ method is worse than that of the DCT method, however the difference is smaller when bit rate of the base layer increases. This is because transform coding compacts energy of images more than pixel based coding. Figure 4 shows a part of a decoded image, pointed out by bold black circles illustrated in Figure 3. We can see detail infor-



(a) Image by the BPQ

(b) Image by the DCT

Fig. 4. Decoded images of the BPQ and DCT methods

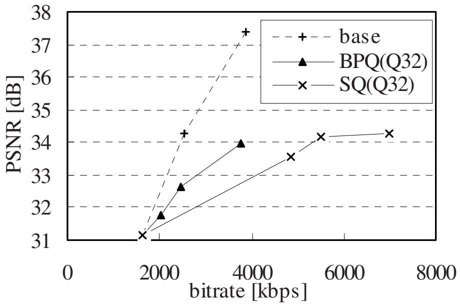


Fig. 5. PSNR of the BPQ and SQ methods

mation, such as the granularity for the grass and the sharpness of the edge for boots, is much more improved, in the image by the BPQ method than the DCT method from the base image, whereas PSNR of the BPQ is almost the same as the DCT. We can conclude the proposed method has different type of distortion feature, which provides better subjective quality than the conventional DCT based method.

4.2 BPQ and Simple Scalar Quantization

The BPQ is compared with the simple scalar quantization (SQ) in terms of PSNR. In the simple scalar quantization, the quantized value "*sq*" is calculated in the encoder side from the residual data "*res*" in the enhancement layer as follows.

$$\begin{array}{ll} sq = (res + 127)/scale + 1 & for\ res >= -127 \\ sq = 0 & for\ res < -127 \end{array} \quad (5)$$

Figure 5 shows the PSNR of the BPQ and the SQ. We can see that coding efficiency of the BPQ is much better than that of the SQ. This is because in the BPQ the larger difference has priority over the lower one to be output.

4.3 Comparison with Other Translation and Entropy Coding Methods

Table 2 lists the average number of bits for the BPQ data. Qx_Rz denotes the data with the BPQ method whose quantization parameter is Rz, when QP of the base images is set to x. The number from other translation method and entropy coding methods are also listed. JPEG-LS was also applied for entropy coding for comparison with AVC based methods. We can see that the translation method LC is always better than MC, and AVC-A achieves the best coding efficiency in most cases except the case of the quantization parameter of the base images is 28 and that of the BPQ is $R1$. From this result, we understand that the efficient translation method and efficient entropy coding method of the BPQ are quite essential for improving coding efficiency.

Table 2. Average number of bits for the BPQ data for other translations (x10^{-5} bpp)

	Translation	JPEG-LS	AVC-I	AVC-P1	AVC-P5	AVC-A
Q28_R1	LC	3.27	3.37	3.31	3.42	3.27
	MC	4.34	4.6	4.92	4.77	4.58
Q32_R1	LC	16.14	14.3	14.36	14.55	13.92
	MC	25.65	22.7	21.01	19.89	19.88
Q28_R3	LC	43.8	37.04	37.62	36.37	36.09
	MC	82.99	71.02	57.72	54.57	54.57
Q32_R3	LC	90.81	77.91	72.93	70.56	70.35
	MC	166.45	144.62	113.1	107.38	107.38

4.4 Comparison with Other Motion Compensation Accuracy for BPQ

Table 3 lists the average number of bits for the BPQ data by several motion compensation methods of the enhancement layer. Other coding process is the same as AVC-A. AVC-A4 uses 1/4 pel MC, which is the same as AVC-A. And AVC-A2 and AVC-A2+ use 1/2 pel MC. The difference of AVC-A2 and AVC-A2+ is the filtering coefficients for generating images at half pel positions. AVC-A2 uses the same coefficients as used for H.264/AVC, and AVC-A2+ uses the

Table 3. Average number of bits for the BPQ data for other mc accuracy (x10^{-5} bpp)

	AVC-A4	AVC-A2	AVC-A2+	AVC-A1	AVC-AA
R2	28.42	27.86	27.86	27.86	27.78
R3	70.35	70.46	70.57	70.64	69.77

same coefficients as used for MPEG-2. AVC-A1 uses integer pel MC. In AVC-AA, type of MC is adaptively selected from AVC-A4, AVC-A2, AVC-A2+, or AVC-A1. Note that when the accuracy of motion is increased, the number of bits for motion vector is generally increased. These results are when the translation method is LC and the quantization parameter of the base layer is 32. We can see that the better type of motion compensation is changed for the range of bit positions of the BPQ, and AVC-AA achieves almost 1-2% bit-rate reduction from AVC-A4. From this result, we understand that the efficient motion compensation method for the BPQ is also essential for improving coding efficiency.

5 Summary

We propose a novel residual coding method in the enhancement layer for scalable video coding. The proposed method uses the bit position quantization (BPQ). The BPQ is developed for representing detail information more efficiently than the quantization in DCT domain. We demonstrate that the BPQ improves more detail information subjectively than the DCT based layered coding, whereas PSNR is almost the same. And we show that LSB centric translation method and AVC based lossless coding method achieve high coding efficiency of the BPQ data. The future works in this study are more efficient entropy coding method for the BPQ data and more effective way of generating the lost high frequency information.

References

1. Sekiguchi, S. and Etoh, M.: Edge-Based Layered Coding of Images. PCS 2001, (2001) 179–182
2. Sun, S. and Lei, S.: On study of a Lossless Video Coding Algorithm Based on H.26L Tools. SPIE Electronic Imaging 2003, 5022–121, (2003)

A Fast Full Search Algorithm for Multiple Reference Images

Hyun-Soo Kang[1], Si-Woong Lee[2], Kook-Yeol Yoo[3], and Jae-Gark Choi[4]

[1] Graduate School of AIM, Chung-Ang University,
221, Heuksuk-dong, Dongjak-ku, Seoul, 156-070, Korea
hskang@cau.ac.kr

[2] Div. of Info. Comm. and Computer Eng., Hanbat National University,
San 16-1, Dukmyung-Dong, Yusong-Gu, Taejeon, 305-719, Korea
swlee69@hanbat.ac.kr

[3] Dept. of Info. and Comm. Eng., YeungNam University,
214-1, Dae-dong, Gyeongsan-si, Gyeongsanbuk-do, 712-749, Korea
kyoo@yu.ac.kr

[4] Dept. of Computer Engineering, Dongeui University,
Gaya-dong, Busanjin-gu, Busan, Korea
cjg@deu.ac.kr

Abstract. This paper presents a fast full search algorithm for motion estimation. The proposed method is an extended version of the rate constrained successive elimination algorithm (RSEA) for multiple reference frame applications. We will show that motion estimation for the reference images temporally preceding the first reference image can be less intensive in computation compared with that for the first reference image. For computational reduction, we will drive a new condition to lead the smaller number of candidate blocks for the best matched block. Simulation results explain that our method reduces computation complexity although it has the same quality as RSEA.

1 Introduction

Motion estimation (ME) have been widely adopted in video systems, since ME is very effective to exploit temporal redundancy of video signals. There is still a lot of need for the methods that can find out motion vectors more accurately and faster. Of ME algorithms, full search algorithm (FSA) yields the optimal motion vectors but requires much computation. To relieve the computational problem, there have been many algorithms such as 2-D logarithmic search algorithm, three step search algorithm, conjugate direct search algorithm, cross search algorithm, four step search algorithm, and diamond search algorithm [1,2,3,4].

Meanwhile, there have been some works to speed up FSA itself without deterioration of the motion estimation error of FSA. The representative works were PDE (partial difference elimination algorithm), SEA (successive elimination algorithm), MSEA (multi-level SEA) and so on. Among them, we would like to pay attention to SEA where a test is performed for whether a search point can

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 615–621, 2004.

be or not a candidate of the optimal vector and thereafter the search points that fail in the test are excluded from the set of candidates for the optimal vector and they are not proceeded further [5]. MSEA can be considered as a generalized version of SEA [7][8]. It hierarchically applies the test done in SEA, varying the resolution of blocks from low resolution to high resolution. And a rate-constrained SEA algorithm was introduced in [6] where the cost function considers both of the sum of absolute difference (SAD) and the number of bits used in motion vectors.

In this paper, we introduce a new method of SEA effectively applicable to multi-reference frame applications such as H.264 [9]. Our method is not for the immediately previous image of a current image, namely the first reference image, but for the other reference images, namely second reference image, third one, and so on. We show that the computation complexity for motion estimation process for the reference images following the first reference image can be reduced using the relation between the first reference image and the other reference images.

2 Background

Prior to explaining our method, we need to introduce a conventional fast algorithm, SEA, for motion estimation. We consider that the size of a block is $N \times N$, the size of the search window is $(2M + 1) \times (2M + 1)$, and $f(p, q, t)$ represents the pixel value at position (p, q) in frame t. Then a block is expressed by As described in [5], the constraint in SEA for relieving the search process is as follows:

$$||\mathbf{M}| - |\mathbf{R}|| \leq SAD(m, n), \tag{1}$$

where $\mathbf{R}$ and $\mathbf{M}$ denote the N^2 dimensional column vectors corresponding to a current block and a reference block, whose elements are $r_{iN+j} = f(p+i, q+j, t)$ for $0 \leq i, j < N$ and with element $m_{iN+j} = f(p + i - x, q + j - y, t - 1)$ for $0 \leq i, j < N$ and $-M < x, y < M$, respectively, and $|\cdot|$ is the sum norm of a vector, for instance,

$$|\mathbf{R}| = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |f(p+i, q+j, t)|, \tag{2}$$

and

$$SAD(m, n) = |\mathbf{R} - \mathbf{O}_{t-1}|, \tag{3}$$

where $\mathbf{O}_{t-1}$ is the N^2 dimensional column vector with element $o_{iN+j} = f(p+i-m, q+j-n, t-1)$. Assuming we have obtained $SAD(m, n)$ for an initial matching candidate block with the motion vector (m, n), Eq.(1) must be satisfied in order that (x, y) is a better matching candidate. Here is the idea of SEA, which is to perform the search process only on those blocks whose sum norms satisfy Eq. (1).

So far we described the fast algorithm considering only distortion regardless of the rate. However, we can minimize the distortion for a given rate constraint.

To do end, in [6], a new cost function is defined with the Lagrangian multiplier as follows:

$$J(x, y, \lambda) = SAD(x, y) + \lambda r(x, y) \tag{4}$$

where λ is the Lagrange multiplier, and (x, y) and $r(x, y)$ are a motion vector and the number of bits associated to the vector. Using the cost function, the condition of Eq. (1) is updated as

$$||\mathbf{M}| - |\mathbf{R}|| + \lambda r(p, q) \leq SAD(m, n) + \lambda r(m, n) \tag{5}$$

3 Proposed Method

3.1 Algorithm Description

In this section, as in the rate constrained SEA [6], which is called RSEA, we will drive an additional inequality to constrain the search process while preserving the optimal motion vector and then describe our algorithm. Since the multiple reference frame approach has been employed by video systems such as H.264 [9], motion estimation methods need to be efficiently extended instead of a straightforward extension. Thus, we introduce a computationally efficient extension of RSEA for two or more reference frames. For simplicity, only two reference frames will be considered. Employing the notation in the previous section, we define a block located at (x, y) in frame $t-2$ as a column vector $\mathbf{P}$ with element $p_{iN+j} = f(p + i - x, q + j - y, t - 2)$ for $0 \leq i, j < N$. Then the constraint on the search process for frame $t-2$ will be derived, while the constraint on frame $t-1$ will follow RSEA.

Prior to explaining the proposed method, we consider the following straightforward implementation of motion estimation for two reference frame: (1) apply RSEA to frame $t-1$, and obtain the optimal motion vector (m^*, n^*) for the frame, and (2) in the same manner, apply it to frame $t-2$, using $J(m^*, n^*, \lambda)$ as an initial minimum cost of the search process for the frame $t-2$. This approach will be called RSEA2 for use in the next section. For convenience of explanation, we assume that the Lagrangian multiplier is given as in R-D optimization of [9]

From Minkowski's inequality, $||\mathbf{A}| - |\mathbf{B}|| \leq |\mathbf{A} - \mathbf{B}|$, we have

$$||\mathbf{P} - \mathbf{M}| - |\mathbf{M} - \mathbf{R}|| \leq |\mathbf{P} - \mathbf{R}|, \tag{6}$$

where $|\mathbf{P} - \mathbf{R}|$ corresponds to the SAD between a current block in frame t and a reference block with displacement of (x, y) in frame $t-2$. Like the derivation in RSEA, let's assume we have obtained $J_2(m, n)$ for an initial matching candidate block with the motion vector (m, n) which is newly defined considering both of reference frames, i.e.,

$$J_2(m, n) = min\{|\mathbf{R} - \mathbf{O}_{t-2}| + \lambda r(m, n, t-2), SAD(m^*, n^*) + \lambda r(m^*, n^*, t-1)\} \tag{7}$$

For a current position (x, y) to be better matching candidate, the following has to be satisfied,

$$|\mathbf{P} - \mathbf{R}| + \lambda r(x, y, t-2) \leq J_2(m, n). \tag{8}$$

Then,

$$||\mathbf{P}-\mathbf{M}|-|\mathbf{M}-\mathbf{R}||+\lambda r(x,y,t-2)\leq J_2(m,n). \tag{9}$$

This is the main result proposed in this paper. $|\mathbf{M}-\mathbf{R}|$ corresponds to the SAD between the current block, $\mathbf{R}$, and the reference block, $\mathbf{M}$, indicated by the motion vector (x,y). And $|\mathbf{P}-\mathbf{M}|$ is the SAD between the blocks, $\mathbf{P}$ and $\mathbf{M}$, located at the same position $(p+i-x,q+j-y)$ in two reference frames at $t-1$ and $t-2$.

Reminding of the procedure of RSEA2, in step (1), if the inequality in Eq. (1) is satisfied, $|\mathbf{M}-\mathbf{R}|$ is computed because the vector is a candidate of the optimal vector. In this case, if $|\mathbf{P}-\mathbf{M}|$ is available by some means, the details of which will be given later, the inequality in Eq. (1) for frame t-2 as well as the inequality in Eq. (9) can be used to constrain the search process for $\mathbf{P}$. This is our key idea where the SAD between the blocks in frame t and frame $t-1$ is used to eliminate the search process for the next previous frame $t-2$, providing step (2) with another additional inequality, Eq. (9). On the other hand, if the inequality in Eq. (1) is not satisfied in step (1), our method will be not applied because $|\mathbf{M}-\mathbf{R}|$ should be additionally calculated which is not on the ordinary process of SEA and RSEA.

Finally, Fig. 1 shows the pseudo-code of our method. It should be noted that our method has computational gain as much as the condition at line 13 is satisfied. That is, the search process for frame $t-1$(line 2 to line 10) is the same as the conventional RSEA, but the one for frame $t-2$(line 11 to line 20) is different. Thus, we have gain in the search process for frame $t-2$. For instance, assume that the condition at line 2 is not satisfied with about 70 percent of blocks and the condition at line 13 is satisfied with half of those blocks. Then, we can save the computation of about 35 percent, $0.7\times 0.5=0.35$, in frame $t-2$. Conclusively, we have gain in the search process for frame $t-2$ depending on the conditions in Eq. (1) and Eq. (9).

3.2 Fast Computation of $|\mathbf{P}-\mathbf{M}|$

$\mathbf{P}$ and $\mathbf{M}$ are associated to blocks located at the same position but in different frames $t-1$ and $t-2$. Thus, $|\mathbf{P}-\mathbf{M}|$ is the SAD between those blocks. To develop its fast computation, suppose that the size of an image is $W\times H$ and the representation of the column vectors, $\mathbf{P}$ and $\mathbf{M}$, is generalized to $\mathbf{P}_{p,q}$ and $\mathbf{M}_{p,q}$, where $\mathbf{P}_{i,j}=\{f(p+i,q+j,t-2),0\leq i,j<N\}$ and $\mathbf{M}_{i,j}=\{f(p+i,q+j,t-1),0\leq i,j<N\}$. Then the computation takes two following steps.

(1) For the whole frames, obtain the absolute difference frame, $d(p,q)=|f(p,q,t-1)-f(p,q,t-2)|$. This requires $W\times H$ sum operations and $W\times H$ absolute operations, or N^2 sum and absolute operations for each block.

(2) In this step, we assume to adopt the unrestricted motion estimation scheme, which removes the need of extraordinary treatment for the blocks near picture boundaries. Then, compute $|\mathbf{P}_{p+1,q}-\mathbf{M}_{p+1,q}|$, using $|\mathbf{P}_{p,q}-\mathbf{M}_{p,q}|$ that have been calculated for the immediately previous search point. $|\mathbf{P}_{p+1,q}-$

$\mathbf{M}_{p+1,q}|$ requires only $2N$ sum operations for $2N$ pixels newly come in $\mathbf{P}_{p+1,q}$ and $\mathbf{M}_{p+1,q}$, and $2N$ subtraction operations for $2N$ pixels gone out from $\mathbf{P}_{p,q}$ and $\mathbf{M}_{p,q}$.

Assuming absolute operation being equivalent to sum operation, on average, the computation overhead for each block OH is

$$OH = 2N^2 + 4N, \tag{10}$$

where $2N^2$ and $4N$ are from step (1) and (2), respectively. Considering block matching at each search point that takes N^2 operations, the computation overhead corresponds to only about two search operations for a large value of N.

```
M_R_available = 0; // flag for check if $|\mathbf{M} - \mathbf{R}|$ is available
if($||\mathbf{M}| - |\mathbf{R}|| > J_2(m,n)$) continue;
else{
  $J_{t-1} = |\mathbf{M} - \mathbf{R}| + \lambda r(x,y,t-1)$;
  M_R_available = 1;
  if($J_{t-1} < J_2(m,n)$){
    $J_2(m,n) = J_{t-1}$;
    $(m,n) = (x,y)$; // update the candidate vector
    Ref_index = 0; // update the reference frame index
}}
if($||\mathbf{P}| - |\mathbf{R}|| + \lambda r(x,y,t-2) > J_2(m,n)$) continue;
else{
  if(M_R_available &&$||\mathbf{P} - \mathbf{R}| - |\mathbf{M} - \mathbf{R}|| + \lambda r(x,y,t-2) > J_2(m,n)$)
  continue;
  else{
    $J_{t-2} = |\mathbf{P} - \mathbf{R}| + +\lambda r(x,y,t-2)$;
    if($J_{t-2} < J_2(m,n)$){
      $J_2(m,n) = J_{t-2}$;
      $(m,n) = (x,y)$;
      Ref_index = 1;
}}}
```

Fig. 1. Pseudo-code of the proposed method

4 Experimental Results

Our algorithm is compared with RSEA2 which is a straightforward extension of RSEA for multiple reference applications. In the experiments, we used seven image sequences with CIF at 30Hz, each of which consists of 100 frames. The block size and the search range are 16 × 16 and ±15, respectively. The number of the reference frames is selected as 2. Adopting the Lagrangian multiplier used in the R-D optimization procedure in H.264, the results of the proposed method

Table 1. Performance of the proposed method and RSEA2

$\lambda =$	Image	ANSP	$ANSP_{t-1}$	$ANSP_{t-2}$		PSNR
$\sqrt{0.85 * 2^{qp/3}}$		of FSA		SEA2	Proposed	
$\lambda = 0$	Coastguard	961	429.6	423.0	122.7	30.41
	Container	961	223.1	221.3	31.4	38.35
	Foreman	961	188.8	179.0	98.3	33.95
	Mobile	961	335.1	334.0	46.6	25.95
	Mot & Dau	961	231.6	229.0	141.4	40.50
	Stefan	961	322.9	311.5	174.6	26.73
	Table tennis	961	469.9	458.5	280.3	30.66
$\sqrt{0.85 * 2^{10/3}}$	Coastguard	961	416.4	408.3	113.1	30.41
	Container	961	190.6	185.1	24.7	38.35
	Foreman	961	173.4	161.5	85.0	33.92
	Mobile	961	325.0	322.6	43.0	25.95
	Mot & Dau	961	206.0	199.7	123.6	40.47
	Stefan	961	307.7	294.5	160.9	26.73
	Table tennis	961	448.7	434.8	263.1	30.65
$\sqrt{0.85 * 2^{20/3}}$	Coastguard	961	384.8	373.1	95.5	30.40
	Container	961	108.8	94.2	9.1	38.34
	Foreman	961	126.5	111.2	51.7	33.83
	Mobile	961	301.4	296.5	37.2	25.94
	Mot & Dau	961	114.9	96.8	49.7	40.39
	Stefan	961	271.9	254.7	139.3	26.72
	Table tennis	961	393.8	372.9	220.1	30.61

and RSEA2 are shown in Table 1, where $ANSP_{t-1}$ and $ANSP_{t-2}$ stand for the average number of search positions per block in the reference frame $t-1$ and $t-2$, respectively, and 'Mot & Dau' in the column of 'Image' means mother and daughter image sequence. Since two methods have the same complexity of motion estimation for frame $t-1$, we should focus on the complexity for frame $t-2$. For comparison, therefore, the column of $ANSP_{t-2}$ has two entries of the proposed method and RSEA2. In $ANSP_{t-2}$, our method saves the computation of 39% to 86% compared with RSEA2. It is interesting that our method is most effective for Container and Mobile sequences which have quite different characteristics. The ANSP for our method should include the block overhead computation. However, since the overhead amounts to just about 2 search points, it can be ignored in consideration of the quantities shown in Table 1. Meanwhile, we can note that ANSP is decreasing as increasing of λ, which means that overhead of our algorithm increases relatively. Thus, our method may be more effective for the case of a small quantization parameter.

5 Conclusion

We proposed a computationally effective extension of SEA for multi-reference frames considering the rate. It saves a considerable amount of computations in motion estimation for the reference frame temporally preceding the first reference frame, while it preserves the same estimation accuracy as FSA. It was realized by adding a new inequality to the inequality of RSEA. Though we dealt with the case of considering two reference frames, it can be easily generalized to the case of more than two reference frames.

Acknowledgements. This research was supported in part by the Ministry of Education, Korea, under the BK21 Project and by the Ministry of Information and Communication, Korea, under ITRC.

References

1. F.Dufaux and F. Moscheni, “Motion estimation techniques for digital TV: A review and a new contribution,” Proc. IEEE, vol, 83, pp. 858–879, June 1995
2. L.M. Po and W. C. Ma, “A novel four-step search algorithm for fast block motion estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 6, pp. 313–317, June 1996
3. L.K.Liu and E. Feig, “A block-based gradient descent search algorithm for block motion estimation in video coding,” IEEE Trans. Circuits Syst. Video Technol., vol. 6, pp. 419–423, Aug. 1996
4. S. Zhu and K.-K. Ma, “A new diamond search algorithm for fast block matching motion estimation,” in Proc. Int. Conf. Inform., Comm., Signal Processing, Singapore, Sept. 9-12, 1997, pp. 292–296
5. W. Li and E. Salari, “Successive elimination algorithm for motion estimation,” IEEE Trans. on Image Processing, vol. 4, no. 1, pp. 105–107, Jan. 1995
6. M. Z. Coban and R. M. Mersereau, “A fast exhaustive search algorithm for rate-constraned motion estimation,” IEEE Trans. on Image Processing, vol. 7, no. 5, May 1998
7. X. Q. Gao, C. J. Duanmu, and C. R. Zou, “A multilevel successive elimination algorithm for block matching motion estimation,” IEEE Trans. on Image Processing, vol. 9, no. 3, pp. 501–504, March 2000
8. J. Y. Lu, K. S. Wu and J. C. Lin, “Fast full search in motion estimation by hierarchical use of Minkowski’s inequality,” Pattern Recognition, vol. 31, no. 7, pp. 945-952, pp. 945–952, 1998
9. ITU-T recommendation H.264 — ISO/IEC 14496-10 AVC, “Draft text of final draft standard for advanced video coding,” Mar. 2003

Preprocessing of Depth and Color Information for Layered Depth Image Coding

Seung-Uk Yoon, Sung-Yeol Kim, and Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea
{suyoon,sykim75,hoyo}@gist.ac.kr

Abstract. The layered depth image (LDI) is a popular approach to represent three-dimensional objects with complex geometry for image-based rendering (IBR). LDI contains several attribute values together with multiple layers at each pixel location. In this paper, we propose an efficient preprocessing algorithm to compress depth and color information of LDI. Considering each depth value as a point in the two-dimensional space, we compute the minimum distance between a straight line passing through the previous two values and the current depth value. Finally, the current attribute value is replaced by the minimum distance. The proposed algorithm reduces the variance of the depth information; therefore, it improves the transform and coding efficiency.

Keywords: Layered depth image, coding, image-based rendering

1 Introduction

Since there have been researches on geometry-based rendering methods, lots of useful modeling and rendering techniques have been developed. However, geometry-based rendering requires elaborate modeling and long processing time. As an attractive alternative to overcome these problems, image-based rendering (IBR) techniques have received much attention. They use two-dimensional (2-D) images as primitives to generate an arbitrary view of the three-dimensional (3-D) scene. IBR techniques require proper computational resources and do not bother from the complexity of 3-D objects in the scene. In addition, it is much easier to acquire a photo or a picture than complex 3-D models of the scene. In spite of these benefits, the amount of data generated from IBR is very huge. Therefore, coding of IBR data is one of the main requirements of IBR techniques.

Various IBR techniques can be classified into three categories based on how much geometry information is used [1], [2]: rendering with no geometry; rendering with implicit geometry; and rendering with explicit geometry. Among the variety of methods, a layered depth image (LDI) [3] is one of the efficient rendering methods for 3-D objects with complex geometries. LDI is contained in rendering with explicit geometry. It represents the current scene using an array of pixels viewed from a single camera location. However, LDI pixel contains not

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 622–629, 2004.

just color values, but also several other attribute values. It consists of color information, depth between the camera and the pixel, and other attributes that support rendering of LDI. Three key characteristics of LDI are: (1)it contains multiple layers at each pixel location, (2)the distribution of pixels in the back layer is sparse, and (3)each pixel has multiple attribute values. Because of these special features, LDI enables us to render of arbitrary views of the scene at new camera positions. Moreover, the rendering operation can be performed quickly with the list-priority algorithm proposed by McMillan [4].

Despite of these benefits, a high resolution LDI contains a huge amount of data [5]. Fig. 1 shows an example. The Cathedral scene occupies 14.1 megabytes (MB), and Stream contains 10.86 MB of data. This means that a single LDI contains a large amount of data, unlike normal 2-D images. If we want to render or represent complex natural scenes with LDI, it requires even higher amount of data. Therefore, it is necessary to compress the LDI data efficiently for the real-time rendering and transmission under a limited bandwidth.

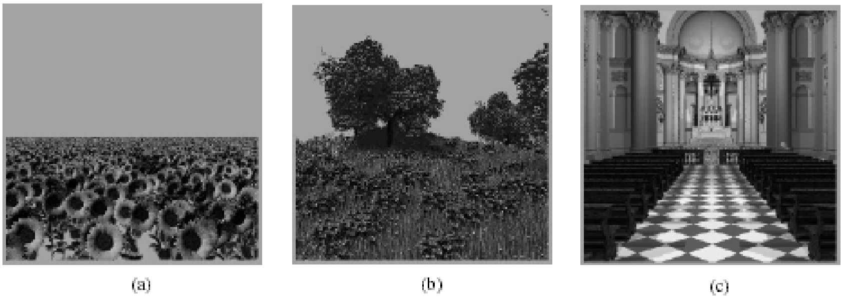


Fig. 1. LDI dataset: (a) Sunflowers, (b) Stream, and (c) Cathedral.

As mentioned in the previous work [5], generic lossless coding tools, such as WinZip, cannot provide the high compression ratio. On the other hand, lossy coding tools, like JPEG-2000 and MPEG-4, guarantee higher coding efficiency, but they cannot be applied directly to compress LDI. Since the density of LDI pixels becomes lower in the back layer, a new algorithm is required to code the LDI data. In order to effectively deal with these features of LDI, a kind of divide and conquer methodologies was proposed by J. Duan *et al.* [5]. In their work, they divide the LDI data into eight components and compress each component image with different techniques. Although their approach provides a high compression ratio with moderate image quality, it does not consider coherency within the component image.

In this paper, we propose a new preprocessing algorithm to improve the transform efficiency. We separate LDI into several component images similarly to the previous work, but our algorithm exploits coherency among pixels within each component image. Considering each pixel as a point in the 2-D space, we compute the minimum distance between a straight line passing through the previous two values and the current depth value. Finally, the current attribute value is replaced by the minimum distance. The proposed algorithm reduces the variance of depth and color information; therefore, it improves the transform and coding efficiency.

The paper is organized as follows. The data structure and previous coding methods of LDI are briefly reviewed in Section 2 and Section 3, respectively. In Section 4, we explain details of our preprocessing algorithm. After experimental results are presented in Section 5, we draw conclusions in Section 6.

2 Layered Depth Image (LDI)

LDI pixels contain depth values along with their colors. In addition, LDI contains potentially multiple depth pixels per pixel location. The farther depth pixels, which are occluded from the LDI center, will act to fill in the disocclusions that occur as the viewpoint moves away from the center. Fig. 2 shows the generation process of LDI. The LDI scene viewed from C_1 is constructed by warping pixels in other camera locations, such as C_2 and C_3.

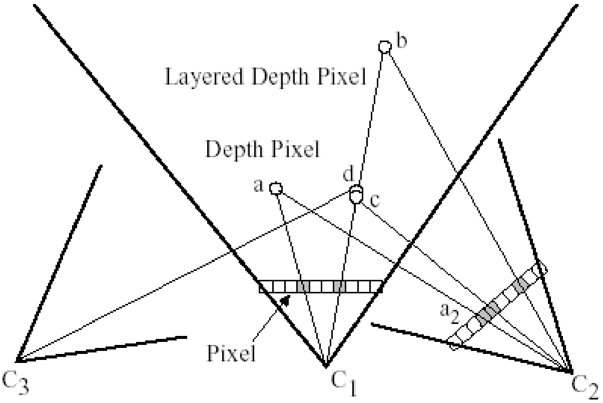


Fig. 2. The generation of the layered depth image.

Unlike the ordinary image consisting of the luminance and chrominance values, each LDI pixel contains 63 bit information [5]: 8 bits each for the R, G and B components, 8 bits for the alpha channel, 20 bits for the depth of the object, and 11 bits for the index into a splat table. The splat table is in turn divided into 5 bits for the distance, 3 bits for the x norm, and 3 bits for the y norm. It is used to support various pixel sizes in rendering of LDI. The overall data structure of the single LDI is shown in Fig. 3.

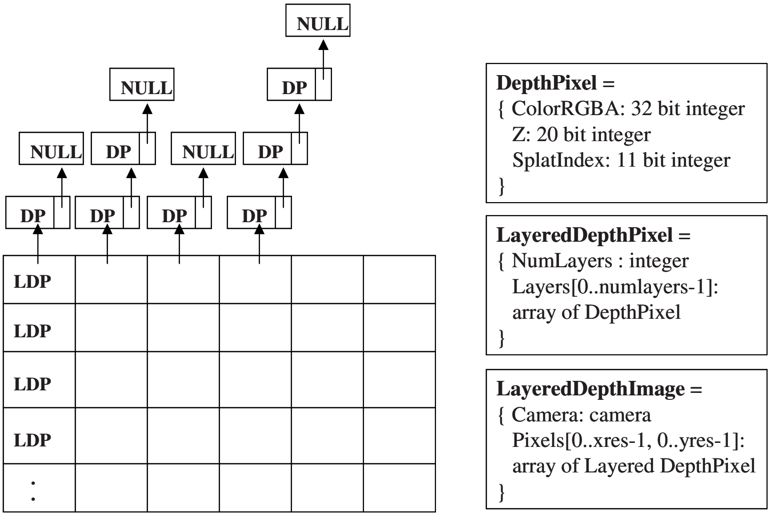


Fig. 3. The structure of LDI.

3 Coding of LDI Data

In the previous work [5], they investigate the compression of the sparse and nonrectangular supported data of LDI. They first record the number of layers (NOL) at each pixel location. The LDI data is then reorganized into a more suitable layout by dividing LDI into layers, each of which contains a mask indicating the existence of pixel in the layer. Each LDI layer is then separated into individual components, such as Y, Cr, Cb, alpha, and depth. Fig. 4 shows eight components of LDI.

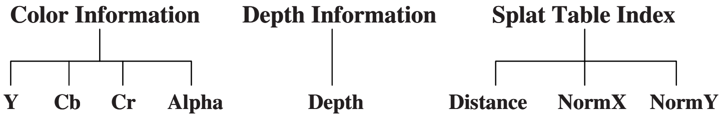


Fig. 4. Layered depth image: component separation.

The component images of each layer are compressed separately. They aggregate the data on the same layer so that the data is more compactly distributed. An arbitrary shape wavelet transform and coding algorithm is used to compress the aggregated data. Finally, the compressed bitstreams of the different layers and components are concatenated to form the compressed LDI bitstream. A practical rate-distortion model is used to optimally allocate bits among all the components.

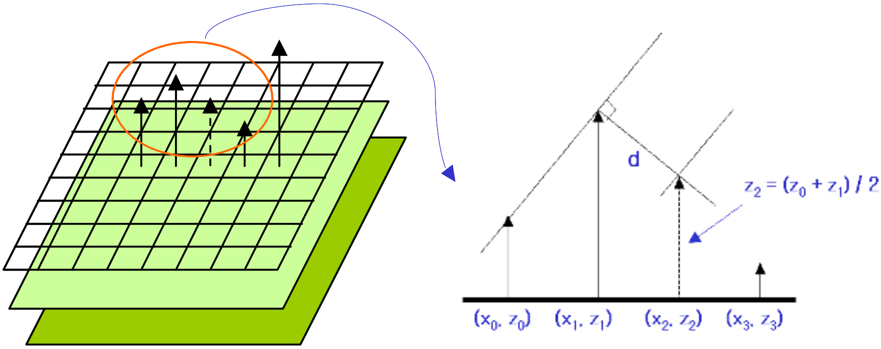


Fig. 5. Calculation of the minimum distance.

4 Preprocessing for LDI Coding

Because of the special data structure of LDI, existing still image compression methods, such as JPEG, cannot be applied directly or are not very efficient. There are three key characteristics of the LDI data. It contains multiple layers at each pixel location; the distribution of pixels in the back layer is sparse; and each pixel has multiple attribute values, including color, depth, and splat table index. In the previous work [5], data aggregation is performed to use these key features of LDI. After aggregating the LDI data, an arbitrary shape wavelet transform and coding algorithm is applied.

In this paper, we propose a new preprocessing algorithm to improve the efficiency of the wavelet transform, which directly affects on the compression ratio of each component image. Thus, our algorithm is performed prior to the wavelet transform. The depth and color information is processed in the same way. Since we observe (x, z) values are changing for the fixed Y-axis, we can consider the one-dimensional (1-D) depth value as the 2-D point. Along the increasing direction of the X-axis, we draw a line passing through two points, and then calculate the Euclidean distance between the line and the current depth value, as illustrated in Fig. 5.

In Fig. 5, the left planes shows the spatial relationship among layers of LDI. The proposed preprocessing method uses correlation among attribute values in the same layer for each component image. In the right figure, the dotted arrow represents the padded depth value at the empty pixel location. We insert the average value of the previous two points into the vacant position. After calculating the minimum distance, we replace the current depth value by the minimum distance. Finally, the inserted average values are removed before the data aggregation. The distance between the line, passing through $A(x_0, z_0)$ and $B(x_1, z_1)$, and the point $C(x_2, z_2)$ is computed by

$$d = \frac{\left|(A-B)^{\perp} \cdot (C-A)\right|}{|A-B|}, \tag{1}$$

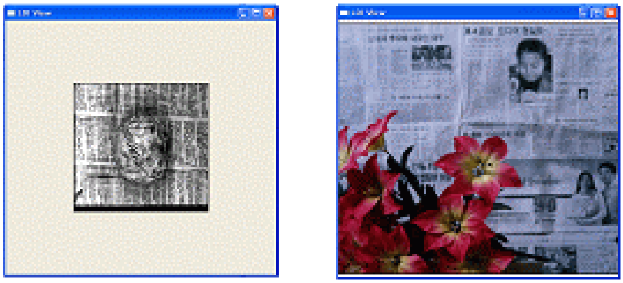
where $A^{\perp}$ is the counterclockwise perpendicular vector to the given A; it means that $(x_0, y_0)^{\perp}$ is $(-y_0, x_0)$.

These procedures are similar to the differential coding method. Instead of calculating the direct difference between two values, we use the distance from the line through the previous two points. Since the direct difference becomes greater in the back layer, the differential methodology is not properly applied to the data structure of LDI. We compare the standard deviation using the differential scheme with the proposed method in our experimental results.

Since each pixel contains the depth and color information at the same location, we can easily compute the minimum distance for color values of Y, Cr, and Cb components; hence, Eq. 1 can be directly reused. In our algorithm, the Euclidean distance is used as the measure for representing the coherency among neighboring depth and color values. This is reasonable because the Euclidean distance is one of the widely used similarity measures for normal 2-D images in general. However, it cannot be applied directly to other attribute values, such as the distance or norms of the splat table index, because their correlations cannot be measured by the Euclidean distance.

5 Experimental Results and Analysis

Efficiency of the proposed preprocessing algorithm is demonstrated with the following experiments. Fig. 6 shows the test data set of LDI scenes. The resolution of Ball LDI is 246 x 246 and that of Flower is 690 x 476. Three layers are used in our experiment for each LDI.



(a) (b)

Fig. 6. Test LDI data set: (a) Ball, (b) Flower.

We calculate the standard deviations for test LDIs to evaluate the performance of the proposed algorithm before and after the preprocessing. As shown in Table 1, the standard deviation of each LDI decreases over 45% after applying the preprocessing algorithm. Especially, Table 1 shows that the standard deviation is reduced much more for Flower LDI. It means that the more pixels, the more reduction occurs because the replaced minimum distance lowers differences among depth values.

Table 1. Standard deviations of depth information

	Ball			Flower		
	Before preprocessing	After preprocessing	Reduction rates	Before preprocessing	After preprocessing	Reduction rates
Layer 1	51.98	27.06	47.94 %	1815.07	364.46	79.92 %
Layer 2	112.80	46.28	58.97 %	3076.95	613.61	80.06 %
Layer 3	154.26	65.66	57.43 %	3740.26	756.03	79.79 %

Table 2 shows the amount of depth information after the wavelet transform and variable length coding. The data size is decreased over 20% because the distribution of depth values is skewed.

Table 2. Amount of depth information of test LDIs [kBytes]

Ball			Flower		
Before preprocessing	After preprocessing	Reduction rates	Before preprocessing	After preprocessing	Reduction rates
96.00	75.70	21.15 %	510.25	405.33	20.56 %

Finally, we compare our algorithm with the differential coding method in terms of the standard deviation. Table 3 shows that the proposed scheme provides higher reduction ratio, because direct differences among depth values become greater than the minimum distance, especially in the back layer. Therefore, the proposed preprocessing algorithm further reduces the variance of depth and color information of LDI.

Table 3. Comparison between the differential technique and the proposed algorithm

	Ball			Flower		
	Original	Differential	Proposed	Original	Differential	Proposed
Layer 1	51.98	21.05	27.06	1815.07	399.82	364.46
Layer 2	112.30	47.24	46.28	3076.95	924.50	613.61
Layer 3	154.26	74.57	65.66	3740.26	1459.07	756.03

6 Conclusions

In this paper, we propose an efficient preprocessing algorithm to code depth and color information of layered depth images. We consider each depth value as a 2-D point. After the minimum Euclidean distance between a line and the current point is calculated, the current depth value is replaced by the minimum distance. Since the previous approach does not consider coherency among neighboring pixels, we focus on using correlations of depth and color values within each component image. Experimental results demonstrate that the proposed algorithm reduces the variance of depth and color information. Therefore, the transform efficiency was improved and the amount of data was reduced.

Acknowledgements. This work was supported in part by Gwangju Institute of Science and Technology (GIST), in part by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC) at GIST, and in part by the Ministry of Education (MOE) through the Brain Korea 21 (BK21) project.

References

1. Shum, H., Kang, S.: A Review of Image-based Rendering Techniques. IEEE/SPIE Visual Communications and Image Processing (VCIP), June (2000) 2–13
2. Shum, H., Kang, S., Chan, S.: Survey of Image-Based Representations and Compression Techniques. IEEE Transaction on Circuit and Systems for Video Technology, Vol. 13, No. 11, November (2003) 1020–1037
3. Shade, J., Gotler, S., Szeliski, R.: Layered Depth Images. ACM SIGGRAPH, July (1998) 291–298
4. McMillan, L.: A List-Priority Rendering Algorithm for Redisplaying Projected Surfaces. University of North Carolina, 95–005 (1995)
5. Duan, J., Li, J.: Compression of the LDI. IEEE Transaction on Image Processing, Vol. 12, No. 3, March (2003) 365–372

Selective Motion Estimation for Fast Video Encoding

Sun Young Lee, Yong Ho Cho, Whoiyul Kim, and Euee S. Jang

College of Information and Communications, Hanyang University,
17 Haengdang-dong, Seongdong-gu, Seoul, 133-791, Korea
{sunny@dmlab.,esjang}@hanyang.ac.kr

Abstract. Motion estimation has been a popular choice in video compression standards such as MPEG-1, MPEG-2, and MPEG-4 to eliminate temporal redundancy of video sequences. The motion estimation and motion compensation (MEMC) part in video encoding represents around 60–80 percent of the total computation, where MEMC is a core part that cannot be avoided for preserving substantial compression efficiency. In this paper, we have presented a method to minimize the use of MEMC, which will further reduce the computational complexity with the reasonable visual quality. The main idea of our proposal is to unselect ME on some frames with very low motion. Experimental results indicated that the proposed coding scheme can reduce about 55–75 percent in computational complexity over the MPEG-4 video reference software.

1 Introduction

Motion compensation based inter-frame coding is a typical video compression method, which is employed in the most video coding standards. Differently from intra-frame coding, motion compensation technique in inter-frame coding can efficiently eliminate temporal redundancy between successive video frames. It, however, increases the encoder complexity due to motion estimation. The motion estimation and motion compensation (MEMC) part in video encoding represents about 60-80 percent of the total computation time [1]. This large computational complexity of MEMC remains as a major obstacle for real-time video encoding [2].

The motion estimation (ME) method as a block matching technique searches the most similar block from the previous video frame, where a video frame is divided into blocks. For the optimal performance in compression efficiency, a full search (FS) can be used in MPEG-4 reference software. FS is a brute-force algorithm, which exhaustively evaluates all possible candidate blocks, requires many computationally intensive loops, and is not a practical solution for real-time services.

Many fast and efficient algorithms of MEMC have been proposed in the past decades, which include the three step search (TSS) [3], the block-based gradient descent search (BBGDS) [4], the diamond search (DS) [5], motion vector field adaptive fast search technique (MVFAST), predictive MVFAST (PMVFAST)

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 630–638, 2004.

[6], and so on. The MPEG-4 part 7 has adopted MVFAST as the core technology for fast motion estimation. PMVFAST is considered as an optional approach for MPEG-4 fast motion estimation. Both MVFAST and PMVFAST are based on diamond search patterns instead of square search patterns and result in fewer search points with similar distortion performance. These techniques can reduce about 50-70 percent in computational complexity.

PMVFAST in [6] showed substantially better performance than FS within a search area set by ±16 pels with no visual distortion of image quality. PMVFAST is an efficient algorithm to reduce complexity, where further complexity reduction is still sought for academia and industry.

As a way to provide acceptable complexity in various cases, we propose a method to minimize the use of MEMC by selectively employing ME per frame. The minimization of MEMC use is possible due to the fact that there is a strong correlation between successive frames, where the collocated previous block may be used for the motion compensation of the current block without further motion estimation. In such a case, we simply skipped the ME, which leads to the further computational complexity reduction while preserving the better visual quality than the case that ME is never employed.

The remainder of this paper is organized as follows. Section 2 describes concept and algorithms of MEMC. Our proposed approach is explained in Section 3. Experimental results are given in Section 4. Finally, we summarize the paper in the last Section.

2 MEMC in MPEG-4

The temporal prediction technique used in MPEG-4 video is based on ME. In the case of no motion or low motion between frames, the encoder efficiently predicts the current frame as a duplicate of the reference frame. The ME is an essential part in eliminating the temporal redundancy between successive video frames, while augmenting the encoder complexity. The encoder of MPEG-4 Visual is composed of texture coding and MEMC. MEMC is the most expensive tool in encoding process because the ME is consuming more than 75 percent of the encoding run time [1][7].

Block matching algorithm (BMA) is one common ME method [8]. A video frame is divided into blocks (8x8 pixels) or macro blocks (MB) (16x16 pixels). Each block of the current frame is compared with all the possible candidate blocks within the search area in a reference frame. In general, the sum of absolute differences (SAD) is used to measure the distortion between the current block and a candidate block. A displacement value with the minimum distortion between the current block and the reference block is selected as a motion vector (MV). The MV coding is performed separately on the horizontal and vertical components.

In order to maximize the coding efficiency, MPEG-4 encodes the differential value between the current MV and the predicted MV formed by the median value of the three vector candidate predictors. These predictors (MV1, MV2, and

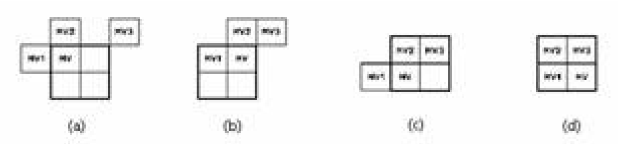


Fig. 1. The candidates for MV predictor

MV3) are derived from the spatial neighborhoods of the current MB, as shown in Fig. 1. The bold square represents the current MB, which is divided into the four blocks (8x8) and the regular squares denote several blocks of neighboring MBs. The selection of the MV predictor candidates depends on the MV mode. In the 1MV mode, the MVs of corresponding 8x8 blocks are set to be the same as the MV of the current MB in Fig. 1 (a). In the case that the 4MV mode is used, the predictor candidates (for each 8x8 block of the MB which is marked MV) shown in Fig. 1 (a) – (d) are used.

2.1 Full Search (FS)

FS is a basic BMA that is currently under use in the MPEG-4 reference software. It is based on square search patterns and searches for the best MV in a coarse-to-fine manner, where all possible displacements are evaluated within the search area of ±16 pels. The following steps describe the FS algorithm.

1) For an MB in the current frame, the best matched MB from the reference frame is found within the search area (±16). In order to find the best matched MB, SAD for the MB (or SAD16) between the founded MB in the reference frame and the MB in the current frame is evaluated.
2) The founded MB is divided into four blocks, of which each BMA is performed within the search area (±2). In this case, SAD for an 8x8 block (or SAD8) is evaluated.
3) Finally, the half-pixel MV of the block, which is with the smaller SAD between the blocks with SAD16 and SAD8, is evaluated,

FS is widely used as a BMA because it is feasible to implement hardware design with regular data flow. It, however, can cause extensive computational complexity by evaluating all possible candidates.

2.2 MVFAST and PMVFAST

The MPEG-4 part 7 has adopted MVFAST as the core technology for fast ME [6][9]. MVFAST makes a significant improvement in regard to both visual quality and encoding speed by the early termination of the search with a diamond pattern compared with the conventional algorithms [10].

MVFAST selects the motion vector predictor (MVP), which is the median vector among the center MV (0, 0) and the MVs of three spatially adjacent blocks. Having MVP selected, MVFAST sets the MVP as the center of search and employs a search with a (large or small) moving diamond. Until the minimum distortion is found at the center of search, the search is to go on moving toward the minimum distortion value.

A diamond search could further assist in initially considering a small set of candidates [6]. The selecting diamond type either a small diamond or a large one is chosen by an initial motion feature of the current block. If the value of the largest MV from three adjacent blocks is smaller than the first threshold T1, A small diamond is used. If it is between the first threshold T1 and the second threshold T2, A large diamond is used instead. Lastly, if it is over the second threshold T2, An additional search could be performed around the center with a small diamond.

PMVFAST is considered as an optional approach for MPEG-4 fast ME [9]. PMVFAST employs fundamentally the search pattern similar to MVFAST [10], evaluates all possible predictors, and uses a method to efficiently decide a diamond pattern [6]. A difference between PMVFAST and MVFAST is the way to select a diamond type, which can lessen encoding time in PMVFAST by using more often a small diamond. As long as the median MV, the median of the three adjacent blocks, is zero and the distortion is relatively larger, a large diamond is used. In other all cases, a small diamond is used.

PMVFAST shows substantially better performance than the other methods with no visual distortion of image quality [6]. Even although PMVFAST is an efficient algorithm to reduce complexity, it may not be appropriate in a low power environment. To reduce further complexity, we propose a method to minimize the use of MEMC by selectively employing ME per frame in the next section.

3 Selective Motion Estimation (SME)

We propose a method of selective motion estimation (SME), where ME is reduced by selectively employing it per frame. Even though fast techniques such as PMVFAST and MVFAST exist, there may be cases when further complexity reduction is required due to the characteristics of input video data. SME provides a way to selectively skip the process of the ME, which leads to further complexity reduction at the cost of marginal visual quality degradation.

In order to skip ME, we need to identify the candidate frames with low motion. We evaluated the motion activity (MA) with the average value of the MVs for a frame using the following equation (1). Fig. 2 shows the average MV value on Akiyo image. To decide which frame is skipped, a threshold T is assigned by a user which is an essential element to control the skip-rate of the ME. If the average MV of the previous frame is below the threshold T in Fig. 2, the ME of the current frame is no longer employed as shown in Fig. 3. Otherwise, PMVFAST is employed for ME. The skip-rate in ME depends on assigning the value of the threshold; in a high value, many frames are skipped and in a low value, few frames are skipped. As a result, users can regulate roughly encoding complexity while preserving acceptable visual quality.

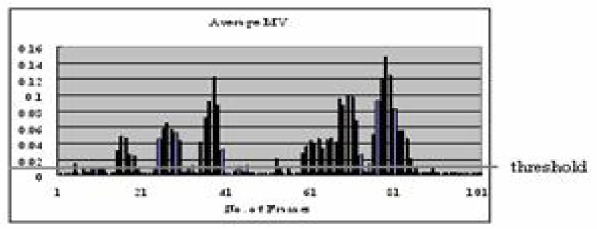


Fig. 2. The average MV for 100 frames on Akiyo sequence

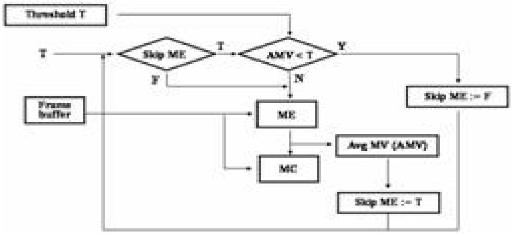


Fig. 3. The structure of the SME

SME has an additional complexity to calculate the average MV per frame and the image quality may not be somewhat as good as the original image quality according to skipping ME. It, however, makes it feasible real-time video encoding service in various power environments by controlling encoding complexity.

$$\mathbf{MA} = \frac{1}{2N}\sum_{i=1}^{N}(|xi| + |yi|),\ where\ N : the\ number\ of\ the\ MBs\ in\ a\ frame \tag{1}$$

3.1 Test Condition

The experiments were done with the Microsoft VM software 2003 edition at various bit rates with no rate control. No additional optimization is applied to reduce complexity when using the codec. A Pentium 4 2.4Hz PC with 512MB RAM and Windows XP as the OS is used for this experiment. In order to evaluate complexity, we measured encoding time with 300 frames. The run-time is averaged over three run times on the test PC. All testing sequences are rectangular videos with the CIF and SIF format.

3.2 Experimental Results

This experimental results show the rate-distortion (RD) performance and the run time analysis of the encoder. We have experimented software with no 4MV in all cases, since MPEG-4 reference software with no 4MV is better than one

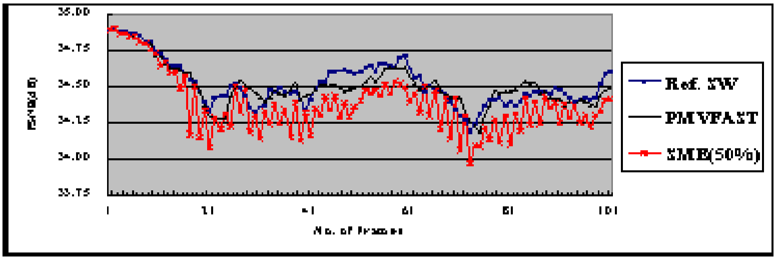


Fig. 4. The rate-distortion curve for 100 frames on Akiyo sequence

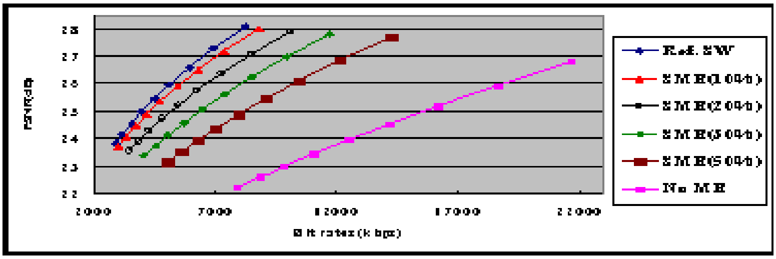


Fig. 5. The rate-distortion curve with SME percent on Stefan sequence

Table 1. Encoding run time of Ref. SW, No ME, and SME 10%, 20%, 30%, and 50%(ms)

	Ref. SW	No ME	SME (10%)	SME (20%)	SME (30%)	SME (50%)
Stefan	52205	11625	49174	45868	42344	31083
	100%	22%	94%	88%	81%	60%

with 4MV in complexity wise with no visual distortion of image quality. SME employs PMVFAST as a ME method which can further reduce the computational complexity compared with the conventional algorithm.

Fig. 4 shows the RD curve of Akiyo sequence for 100 frames respectively. It can be noticed that the image quality of the skipped frames is abruptly fallen down while preserving the reasonable PSNR performance for the whole sequence. In fact, the ME of this image is curtailed around 50 percent (or skip-rate) for reducing complexity. The skip-rate is assigned by a threshold T users decide to set. The RD performance on various skip-rate values is given in Fig. 5. When the skip-rate is higher, the PSNR will be lower with the complexity reduction in Table 1. This implies that SME encoder can manipulate the encoding speed in a flexible way with marginal degradation of image quality.

In Fig. 6 and Fig. 7, we have presented the test results on MPEG-4 reference software, PMVFAST, SME (50% skip), and so on. The PSNR performance of SME on the two sequences shows a marginal degradation of visual quality with improved computational complexity reduction. The encoding time perfor-

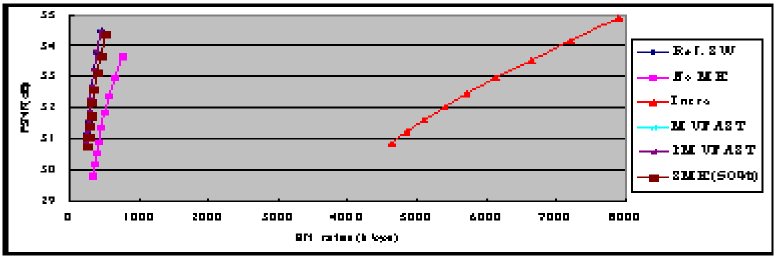


Fig. 6. The rate-distortion curve on Akiyo sequence

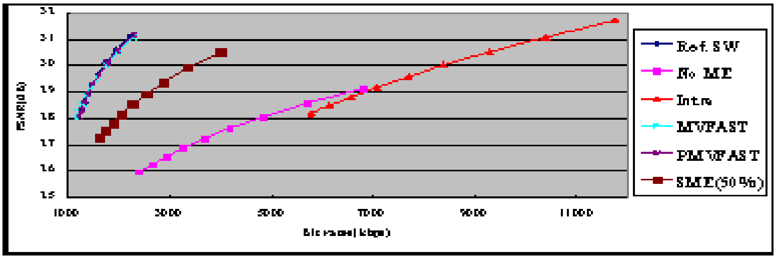
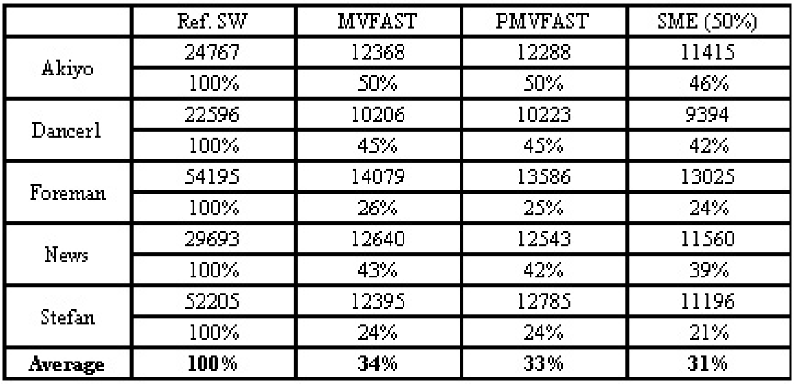


Fig. 7. The rate-distortion curve on Foreman sequence

Table 2. Encoding run time of Ref. SW, Intra, No ME, PMVFAST, and SME 50% (ms)



	Ref. SW	MVFAST	PMVFAST	SME (50%)
Akiyo	24767	12368	12288	11415
	100%	50%	50%	46%
Dancer1	22596	10206	10223	9394
	100%	45%	45%	42%
Foreman	54195	14079	13586	13025
	100%	26%	25%	24%
News	29693	12640	12543	11560
	100%	43%	42%	39%
Stefan	52205	12395	12785	11196
	100%	24%	24%	21%
Average	**100%**	**34%**	**33%**	**31%**

mance is shown in Table 2. PMVFAST showed substantially better performance than the reference in complexity. SME, moreover, can reduce further complexity compared to PMVFAST and make encoding time flexible by using skip-rate. Fig. 8 presents the first skipped frame image on Akiyo, where a very marginal degradation can be confirmed.

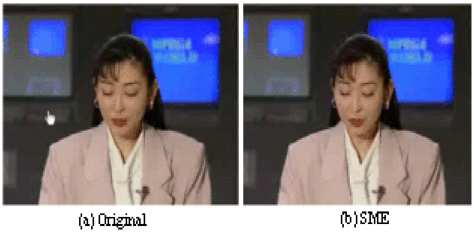


Fig. 8. The first skipped frame of Akiyo sequence; (a) original and (b) SME

4 Conclusions

We presented an efficient method to minimize the use of the MEMC, which lessens computational complexity of ME by selectively utilizing it per frame. Minimizing ME can reduce about 55–75 percent in complexity over the MPEG-4 video reference software and is less complex than PMVFAST, which is an efficient method to show good performance in complexity. A flexible management of encoding speed and visual quality is possible with the proposed method by a threshold. It should be noted that the flexible encoding speed was possible at the cost of marginal visual quality degradation. Our future work will be dedicated to devise methods to enhance the visual quality by using adaptive threshold value or a little ME for the skipped frames while preserving the encoding speed flexibility.

Acknowledgements. This work was supported by the Hanyang University Research Fund.

References

1. W. Zheng, I. Ahmad, and M. L. Liou, “Benchmark the software based MPEG-4 video codec,” International IEEE Conference on Electronics, Circuits and Systems, vol. 1, pp. 289–292, 2001
2. Euee Seon Jang, “Low-complexity MPEG-4 shape encoding towards realtime object-based applications,” ETRI journal, vol. 26, no.2, pp. 122–135, Apr. 2004
3. T. Koga, K. Linuma, A. Lijima, and T. Lshiguro, “Motion-compensated interframe coding for video conderencing,” Proc. NTC81, New Orleans, LA, pp. C9.6.1–9.6.5, 1981
4. L. M. Po and W. C. Ma, “A novel four-step search algorithm for fast block motion estimation,” IEEE Trans. Circuits Syst. Video Tech., vol. 4, pp. 438–442, Aug. 1994
5. S. Zhu and K.-K. Ma, “A new diamond search Algorithm for fast block-matching motion estimation,” IEEE Trans. Image Processing, vol. 9, pp. 287–290, Feb. 2000
6. A. M. Tourapis, O. C. Au, and M. L. Liou, “Predictive Motion Vector Field Adaptive Search Technique (PMVFAST) – Enhancing Block Based Motion Estimation,” proceedings of Visual Communications and Image Processing 2001 (VCIP-2001), San Jose, CA, January 2001

7. H.-C. Chang, Y.-C. Wang, M.-Y. Hsu, and L.-G. Chen, “Efficient algorithms and architectures for MPEG-4 object-based video coding,” IEEE workshop on Signal Processing Systems, pp. 13-22, 2000
8. W. Zheng, I. Ahmad, and M. L. Liou, “Adaptive motion search with elastic diamond for MPEG-4 video coding,” IEEE Conference on Image Processing, 2001. Proceedings 2001, vol. 1, 7-10 pp. 377–380, Oct. 2001
9. ISO/IEC JTC1/SC29/WG11 N3324, “Optimization Model Version 1.0,” Noordwijkerhout, NL, Mar. 2000
10. J. Y. Tham, S. Ranganath, M., Ranganath, A. A. Kassim, “A novel unrestricted center-biased diamond search algorithm for block motion estimation,” IEEE Trans. On Circuits and Systems for Video Technology, vol. 8, no. 4, pp. 369–377, Aug. 1998

An Efficient VLSI Architecture of the Sample Interpolation for MPEG-4 Advanced Simple Profile

Lei Deng, Ming-Zeng Hu, and Zhen-Zhou Ji

Department of Computer Science and Engineering
Harbin Institute of Technology, Harbin 150001, China
ldeng@jdl.ac.cn

Abstract. Sample interpolation which has a computationally expensive finite impulse response (FIR) digital filter is one of the key modules in MPEG-4 Advanced Simple Profile (ASP). Normal FIR architectures have low efficiency on its implementation due to the special input data stream. In this paper, based on a pure systolic FIR, the efficient VLSI architecture for the sample interpolation has been implemented. Experimental result shows that the efficiency of proposed architecture is three times higher than normal ones, and it satisfies the applications such as MPEG-4 ASP.

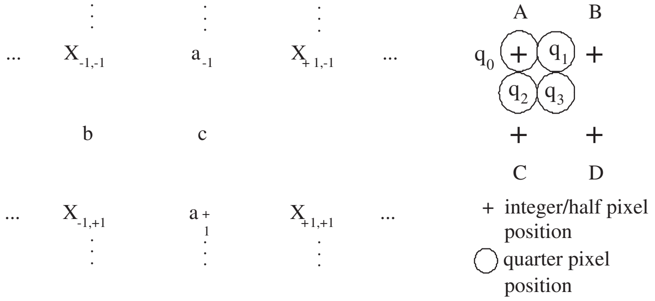
1 Introduction

The MPEG-4 standard [1] is a standardized framework for many multimedia applications, such as teleshopping, teleconferencing, mobile video communication and interactive, because of its high coding efficiency and high error resilience. MPEG-4 ASP is defined in Streaming Video Profile in Amendment of MPEG-4 [4]. Different with MPEG-4 Simple Profile, ASP contains many power tools such as B-frame, global motion compensation, and quarter sample motion compensation (MC) interpolation which are defined for implementations with higher processing capability. But the sample interpolation is one of the critical paths of MPEG-4 ASP decoder, due to its computationally expensive process and special input data streams. Above all, the MC interpolation procedure is described below.

The sample interpolation has two stages. In the first stage, the algorithm calculates the half sample values by an 8-tap FIR filter such as a_i by horizontal filtering (eq.(1)) and b, c by vertical filtering (eq.(2),eq.(3)). The filter coefficients is: $CO1[1..4] = [160, -48, 24, -8]$. Fig. 1a shows these half samples position in the frame, where X represents position of integer sample. After filtering, they are clipped to the range of $[0, 2N\_bit - 1]$, where N_bit is the number of bits per pixel.

$$a_i = \left(\left(\sum_{j=1}^{4} co1[j] \cdot (x_{-j,i} + x_{+j,i}) \right) + 128 - rounding\_control \right) \Big/ 256 \qquad (1)$$

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 639–646, 2004.



1(a) half stage

1(b) quarter stage

Fig. 1. The position of half samples and quarter samples

$$b = \left(\left(\sum_{i=1}^{4} co1[i] \cdot (x_{-1,-i} + x_{-1,i})\right) + 128 - rounding\_control\right) \Big/ 256 \tag{2}$$

$$c = \left(\left(\sum_{i=1}^{4} co1[i] \cdot (a_{-i} + a_{+i})\right) + 128 - rounding\_control\right) \Big/ 256 \tag{3}$$

In the second stage, the quarter sample values are calculated by the bilinear interpolation between integer and the corresponding half sample values, see Fig. 1b which shows the displacement between quarter and half position. According to the different position, the bilinear interpolations are:

$$q0 = A \tag{4}$$

$$q1 = \big(A + B + 1 - rounding\_control\big)/2 \tag{5}$$

$$q2 = \big(A + C + 1 - rounding\_control\big)/2 \tag{6}$$

$$q3 = \big(A + B + C + D + 2 - rounding\_control\big)/4 \tag{7}$$

Before interpolation, the data must be fetched from the out-chip memory. For each block of size $m \times n$ in the reference VOP, a reference block of size $(m + 1) \times (n + 1)$ should be read from the reconstructed reference VOP. Once the reference pixels have been obtained, the next step is to mirror the first three and the last three pixels on each side outwards. And then the 8-tap FIR filter calculates this extended reference block. It is known that the length of the 8-tap FIR filter's response M is 8, and we define N as the length of the input data stream for the 8-tap FIR filter and S as the number of input data streams. In fact, N is the amount of the pixels in a line of the mirrored reference block and S is the amount of the lines in the block. After mirroring N is equal to $n + 1 + 3 + 3$ and S equals $m + 1 + 3 + 3$. Usually n and m take the value of 8, which means an input data stream has only 15 pixels in a line. Consider of the response of the 8-tap FIR filters, the common implementation [2–3] have so low efficient that they can not meet the need of the MPEG-4 ASP decoder. This paper presents a new efficient systolic architecture of FIR filter, and gets an efficient VLSI implementation of quarter sample MC interpolation.

In Section 2, the efficient FIR filter architecture is described in detail. In Section 3, the architecture of quarter sample mode interpolation is presented. Finally, the experimental result and conclusion are given in Sections 4 and 5, respectively.

2 The Efficient FIR Filter Architecture

2.1 A Referenced Pure Systolic FIR

For efficient VLSI implementation, a systolic array of FIR is showed in Fig. 2. In it, the number of PE is M. response h_i is stored in PE_i. In every other cycle, input data stream X sends a pixel which will move from the left to the right and output data stream Y receives a result at the left of the array. There is no global data bus in the Architecture, so this is especially suitable for VLSI implementation. But in the systolic array, only a half of PEs work at the same time, so it can't get high efficient enough for quarter sample process.

Figure 3 is the timing chart of filtering the length-N input data stream by the systolic array. In Fig. 3, the shaded cycle represents that the PE has a valid

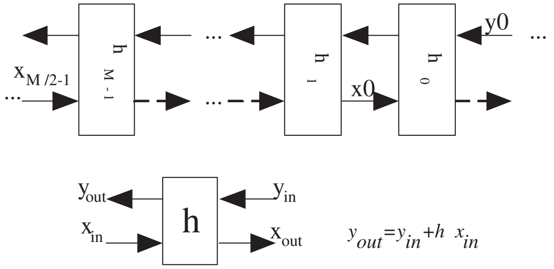


Fig. 2. A pure systolic array of FIR

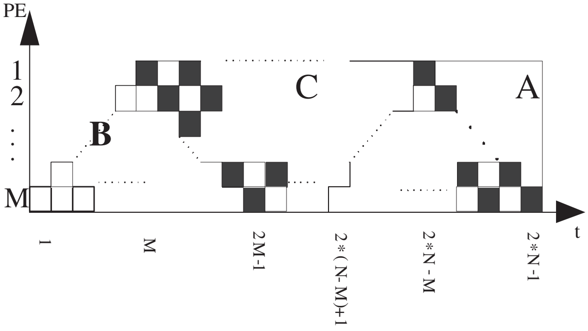


Fig. 3. The timing chart

calculation at the time, and oppositely the white cycle represents an invalid calculation. Since the input data stream X sends pixels in every other cycle, each PE only has a valid calculation in every two cycles and valid work and invalid work are alternate. So the efficiency of the array can be written as:

$$E = \frac{M \times (N - M + 1)}{M \times (2 \times N - 1)} = \frac{1}{2} - \frac{M - 3/2}{2 \times N - 1} \tag{8}$$

In quarter sample interpolation, usually $N = 15$ and $M = 8$, so E is approximately 27.6%. From eq.(8), firstly it can be seen that E is always less than1/2. Secondly there is an inverse relation between M and E, and a direct relation between N and E. That means the array efficiency E will higher if the response length of the FIR filter is smaller and the input data stream is longer. Thirdly in our application N is more close to M, so compare with valid cycles it makes the array have too many cycles staying in startup state. Although the array is very suitable to VLSI, it is not much efficient for MC sample interpolation. We need both the efficiency and the nice VLSI implementation, so this paper improves upon the systolic architecture of Fig. 2 below.

2.2 An Improvement of Pure Systolic FIR

If defining E in another way as follow:

$$E = 1 - \frac{A + B + C}{M \times (2 \times N - 1)} \tag{9}$$

Where A, B, C is the quantity of invalid calculations in different areas of Fig.3 and

$$A = \frac{M \times (M - 1)}{2} \qquad B = \frac{3M \times (M - 1)}{2} \qquad C = M \times (N - M) \tag{10}$$

We can see that A, B and C are side effect on the efficiency of the array. A and B are the invalid cycles needed by the array initialization between two different input data streams, and they are only decided by M. Since one input data stream should have one initial process and many streams in a reference block, there have lots of initializations during the sample interpolation of a reference block. So A and B may decrease E sharply. C are the invalid cycles after the initialization, since only a half of PEs work at the same time and the valid cycle and the invalid cycle are alternate after the initialing of the array. C also lowers the efficiency of the array.

In order to increase the pipeline efficiency, a new efficient FIR systolic architecture (Fig. 4) is gotten from the below analysis. Two modifications are made in the architecture of Fig. 2. Firstly, input data stream X is instead of X, and the data arrangement of X is $\ldots x_2^{'}x_2x_1^{'}x_1x_0^{'}x_0$, where x_i and $x_i^{'}$ are come from two original input data stream. X sends pixels in each cycle, and correspondingly the output can be gotten in each cycle too. All PEs may work in each cycle, the format of an output data stream is $y_0^{'}y_0y_1^{'}y_1y_2^{'}y_2 \ldots$, where y_i and $y_i^{'}$ are in two

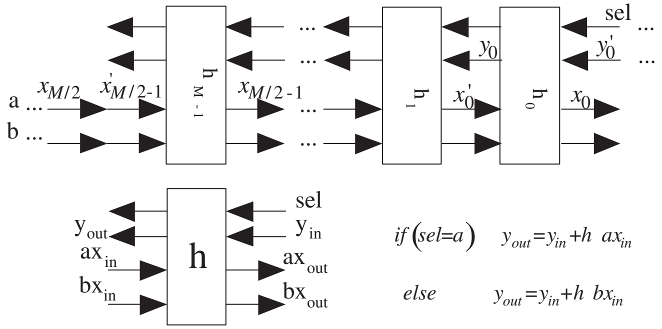


Fig. 4. The new efficient FIR systolic architecture

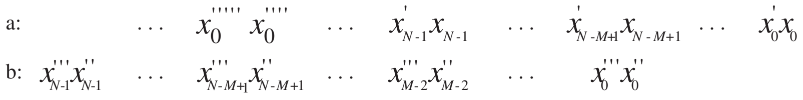

a: ... x_0''''' x_0'''' ... x'_{N-1} x_{N-1} ... x'_{N-M+1} x_{N-M+1} ... x'_0 x_0
b: x'''_{N-1} x''_{N-1} ... x'''_{N-M+1} x''_{N-M+1} ... x'''_{M-2} x''_{M-2} ... x'''_0 x''_0

Fig. 5. The timing relation between two data ports

original output data streams. Invalid calculations represented by C are avoided. Secondly, another data bus and a control signal are added in the array in order to make more parallel between lines of reference block.

The timing between two data ports is shown in Fig. 5. Data x_0 is sent to port a firstly, and the data $x_0^{''}$ is sent to port b at the same cycle when x_{N-M+1} is being inputted to a. If all data in a stream are consumed, the next stream for port a waits or inputs immediately. If N is larger than $2M-3$ the next stream begin to send data to port a when $x_{N-M+1}^{''}$ is being sent to port b, else the stream starts to input when $x_{M-2}^{''}$ is being inputted to b. the input of the next stream for port b is just like the case of port a. If N is larger than $2M-3$, the cycle for the first data of the input data stream X_i for port a is:

$$t_i = 4 \times i \times \left(N - M + 1\right) \qquad i \geq 0 \tag{11}$$

and for port b is:

$$t_i^{'} = 2 \times \left(2 \times i + 1\right) \times \left(N - M + 1\right) \qquad i \geq 0 \tag{12}$$

Under this condition, after the startup state, all PEs are busy till achieving the sample interpolation of the reference block. There have space between two input data streams of the same port, so the throughput of the array is smaller than that of the data ports. If N is no more than $2M-3$, the cycle for the first data of the input data stream X_i for port a is:

$$t_i = 2 \times i \times N \qquad i \geq 0 \tag{13}$$

and for port b is:

$$t_i^{'} = 2 \times i \times N + 2 \times \left(N - M + 1\right) \qquad i \geq 0 \tag{14}$$

The data is sent to port a or port b continually, and PEs have idle cycles between two input data streams. So the throughput of the data ports is smaller than that of the array. The control signal is used to choose a data from port a or port b as the processing data of the PE. The selection of the signal is shown as Eq. (15).

$$sel = \begin{cases} a & t_i + M \leq t < t_i^{'} + M \\ b & t_i^{'} + M \leq t < t_{i+1} + M \end{cases} \tag{15}$$

The efficiency of Fig. 4 is:

$$E = 1 - \frac{2M - 2 + p}{t + 2N - 1} \tag{16}$$

where

$$t = \begin{cases} t^{'}_{\lceil S/2 \rceil /2} & \lceil S/2 \rceil \text{is even} \\ t_{\lceil S/2 \rceil /2+1} & \lceil S/2 \rceil \text{is odd} \end{cases} \tag{17}$$

$$p = \begin{cases} 0 & N > 2M - 3 \\ (\lceil S/2 - 1 \rceil) \times (2M - N - 2) & N \leq 2M - 3 \end{cases} \tag{18}$$

And S is the number of input data streams. In quarter sample mode interpolation, usually $S \leq 12$, so the maximum of E is about 90.6%.

3 The Architecture of Sample Interpolation

The implementation architecture of the quarter sample MC interpolation is shown in Fig. 6. Before the interpolation of a size $m \times n$ block, a reference block of size $(m+1) \times (n+1)$ is read from the reconstructed and padded reference VOP and putted in $FIFO\_a$ and $FIFO\_b$ after mirroring. Then FIR calculates the half sample value (sample1 in Fig. 6) in each cycle, and output the integer pixel of current position or right of current position (sample2 in Fig. 6). Sample1 and sample2 are used in half sample interpolation to produce quarter sample values or horizontal filtering results. If horizontal filtering results, it should be pass through the pipeline again in order to get the final results (quarter sample values).

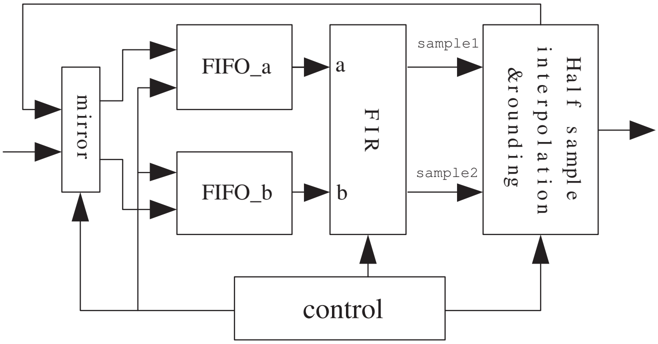


Fig. 6. The architecture of the quarter-pel interpolation

Table 1. The FIR performance of Fig.2 and Fig.3

	area(um_2)	CLK(ns)
FIR1 in fig.2	60647.067132	4.81
FIR2 in fig.3	89341.000000	4.87

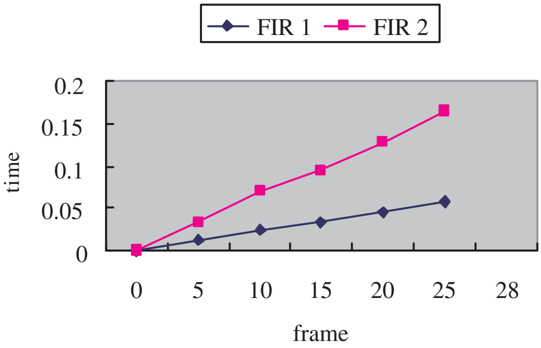


Fig. 7. Cycles of FIR1 and FIR2

4 Experimental Result

After synthesizing using TSMC 0.25um CMOS 1P5M process, the architecture works at 200MHz clock rate, and the total area is about 428749.312000 um^2 including memory component. The FIR performance of Fig. 2 and Fig. 3 shows in table. 1. Quarter sample mode interpolation is time consuming operation in MPEG-4 decode application. The aim of high efficiency is to shorten the total cycles of calculation on this process with low area consumption. After Simulating $FIR1$ and $FIR2$ under two video sequences in Verilog-XL, the run time is shown in Fig. 7. In Fig. 7, the cycles used by $FIR2$'s are about one third of the $FIR1$. This verifies the above conclusions which efficiency of $FIR1$ and $FIR2$ is 27.6% and 90.6%.

5 Conclusions

This paper describes an efficient systolic FIR architecture for the VLSI implementation of the sample interpolation. The architecture works at 200MHz clock rate, and the total area of it is about 428749 um^2 including memory component. The maximum efficiency of the FIR is 90.6%, which is three times of Fig. 2. Area of it is about 89341 um^2. The design satisfies the MPEG-4 decoder applications.

References

1. ISO/IEC 14496-2 1999/Amd.1 2000: Coding of Audio-Visual Objects-Part 2 Visual, Amendment 1 Visual Extensions. Maui, Dec. 1999
2. John Bombardieri: systolic pipeline architectures for symmetric convolutions. ieee transactions on signal processing. Vo.40, no.5, May 1992
3. Shang Yong, Wu Shunjun: Design of parallel adaptive FIR filters. IEEE APCCAS'98, Chiangmai, Thailand:1998, 81–84
4. ISO/IEC 14496-2 1999/FDAM4 Amendment 4 Streaming Video Profile. La Baule, Oct. 2000

Performance Improvement of Vector Quantization by Using Threshold

Hung-Yi Chang[1], Pi-Chung Wang[2], Rong-Chang Chen[3], and Shuo-Cheng Hu[4]

[1] Department of Information Management
I-Shou University, Ta-Hsu Hsiang, Kaohsiung County, Taiwan 840, R.O.C.
leorean@isu.edu.tw
[2] Institute of Computer Science and Information Technology
[3] Department of Logistics Engineering and Management
National Taichung Institute of Technology, Taichung, Taiwan 404, R.O.C.
{abu,rcchens}@ntit.edu.tw
[4] Department of Information Management
Ming-Hsin University of Science and Technology, Hsinchu, 304 Taiwan, R.O.C.
schu@mis.must.edu.tw

Abstract. Vector quantization (VQ) is an elementary technique for image compression. However, the complexity of searching the nearest codeword in a codebook is time-consuming. In this work, we improve the performance of VQ by adopting the concept of *THRESHOLD*. Our concept utilizes the positional information to represent the geometric relation within codewords. With the new concept, the lookup procedure only need to calculate Euclidean distance for codewords which are within the threshold, thus sifts candidate codewords easily. Our scheme is simple and suitable for hardware implementation. Moreover, the scheme is a plug-in which can cooperate with existing schemes to further fasten search speed. The effectiveness of the proposed scheme is further demonstrated through experiments. In the experimental results, the proposed scheme can reduce 64% computation with only an extra storage of 512 bytes.

1 Introduction

Currently, images have been widely used in computer communications. The sizes of images are usually huge and need to be compressed efficiently for storage and transmission. Vector quantization (VQ) is an important technique for image compression, and has been proven to be simple and efficient [1]. VQ can be defined as a mapping from k-dimensional Euclidean space into a finite subset C of R^k. The finite set C is known as the *codebook* and $C = \{c_i | i = 1, 2, \ldots, N\}$, where c_i is a *codeword* and N is the codebook size.

To compress an image, VQ comprises two functions: an encoder and a decoder. The VQ encoder first divides the image into $N_w \times N_h$ blocks (or vectors). Let the block size be k ($k = w \times h$), then each block is a k-dimensional vector. VQ selects an appropriate codeword $c_q = [c_{q(0)}, c_{q(1)}, \ldots, c_{q(k-1)}]$ for each image

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 647–654, 2004.

vector $x = [x_{(0)}, x_{(1)}, \ldots, x_{(k-1)}]$ such that the distance between x and c_q is the smallest, where c_q is the closest codeword of x and $c_{q(j)}$ denotes the jth-dimensional value of the codeword c_q. The distortion between the image vector x and each codeword c_i is measured by their *squared Euclidean distance*, i.e.,

$$d(x, c_i) = \|x - c_i\|^2 = \sum_{j=0}^{k-1} [x_{(j)} - c_{i(j)}]^2. \tag{1}$$

After the selection of the closest codeword, VQ replaces the vector x by the *index* q of c_q. The VQ decoder has the same codebook as that of the encoder. For each index, VQ decoder can easily fetch its corresponding codeword, and piece them together into the decoded image.

The codebook search is one of the major bottlenecks in VQ. From equation (1), the calculation of the squared Euclidean distance needs k subtractions and k multiplications to derive k $[x_{(j)} - c_{i(j)}]^2$s. Since the multiplication is a complex operation, it leads to the increase of the degree of the total computational complexity of equation (1). Therefore, speeding up the calculation of the squared Euclidean distance is a major hurdle.

Many methods have been proposed recently to shorten VQ encoding time [2,3,4,5,6]. The simplest one among them is the *look-up table* (LUT) method [4]. It suggests that the results of the $[x_{(j)} - c_{i(j)}]^2$s for all possible x_j and y_{ij} should be pre-computed first, and then stored into a huge matrix, the LUT. Suppose the values of $x_{(j)}$ and $c_{i(j)}$ are within $[0, m-1]$. Then the size of matrix L should be $m \times m$ and

$$L = \begin{bmatrix} 0 & 1^2 & \cdots & (m-1)^2 \\ 1^2 & 0 & \cdots & (m-2)^2 \\ \vdots & \vdots & \ddots & \vdots \\ (m-1)^2 & (m-2)^2 & \cdots & 0 \end{bmatrix} \tag{2}$$

Given any $x_{(j)}$ and $c_{i(j)}$, we can get the square of their difference directly from $L[x_{(j)}, c_{i(j)}]$. Therefore, equation (1) could be rewritten as follows:

$$d(x, c_i) = \sum_{j=0}^{k-1} [x_{(j)} - c_{i(j)}]^2 = \sum_{j=0}^{k-1} L[x_{(j)}, c_{i(j)}]. \tag{3}$$

LUT can be employed to avoid the subtractions and the multiplications in equation (1). Hence, it is an efficient method.

As discussed above, the designed criteria of LUT-based schemes emphasize computation speed, table storage and image quality. However, the number of calculated codewords has not mentioned since these schemes did not utilize the geometrical information implied in the codewords. In this work, we introduce a new concept of *THRESHOLD* which reduces the number of the computed codewords. The new scheme uses two integers to represent the geometric relation within codewords. Accordingly, the search procedure could refer the integers to sift candidate codewords easily. The scheme might also cooperate with existing schemes to fasten search speed. The proposed scheme is simple and suitable for hardware implementation. Through experiments, the proposed scheme can reduce 64% computation with only 512-byte storage. The rest of this paper is organized as follows. The proposed scheme is presented in Section 2. Section 3 addresses the performance evaluation. Section 4 concludes the work.

2 The Proposed Scheme

The concept of *THRESHOLD* is based on the assumption that the distance of each dimension between the tested vector and the selected codeword is small. Since VQ selects the codeword with the smallest Euclidean distance to the tested vector, the selected codeword can be treated as the nearest point in the k-dimensional space. In the ideal case (with a well-trained codebook), the tested vector and the selected codeword should be relatively close in each dimension. Hence it is possible to filter out unfeasible codewords by referring the positional information.

Figure 1 presents the distribution of the maximal one-dimensional distance $\max_{j=0}^{k-1} |x_{(j)} - c_{M(j)}|$ between the tested vectors and their matched codewords. The codebook is trained according to the image "Lena", and 5 images are compressed by full search. PSNR (*peak signal-to-noise ratio*) is defined as $PSNR = 10 \cdot log_{10}(255^2/MSE)$ dB. Here the MSE (*mean-square error*) is defined as $MSE = (1/H) \times (1/W) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} [\alpha_{(i,j)} - \beta_{(i,j)}]^2$ for an $H \times W$ image, where $\alpha_{(i,j)}$ and $\beta_{(i,j)}$ denote the original and quantized gray level of pixel (i, j) in the image, respectively. A larger PSNR value has been proven to have preserved the original image quality better. For the compressed images with better quality, including "Lena" and "Zelda", about 98% of their maximal one-dimensional distances are less than 32. However, the ratio is reduced to 91% $\sim$ 96% for the other images since their quality of compression is also decreased.

We use a two-dimensional case as an example to explain our idea. There are two codewords, C_1 (3, 1) and C_2 (2, 3), as shown in Fig. 2. To calculate the nearest codeword for the tested vector, V_1 (1, 2), the squared Euclidean distances to C_1 and C_2 are 5 and 2, respectively. Hence C_2 should be chosen as the result. Also, C_1 is selected for V_2.

To utilize the observation presented in Fig. 1, we set a range for the vertical axis to prune the codewords. As shown in Fig. 3(a), only the codewords whose distance in the vertical axis is smaller than 2 are considered.

If we consider only the distances in the vertical axis rather than the Euclidean distances, the computation procedure is changed as follows. Assuming the maximal distance in the vertical axis is set to 1, i.e., only the codewords whose distance in the vertical axis, D, is smaller than or equal to 1 is considered in the calculation of the Euclidean distance. For the example in Fig. 3(a), C_1 and C_2 are calculated for the Euclidean distance to V_1. However, only C_1 is chosen for V_2 and the calculation for C_2 could be avoided. In some conditions, there is no candidate for the tested vector. As a result, each codeword has to be

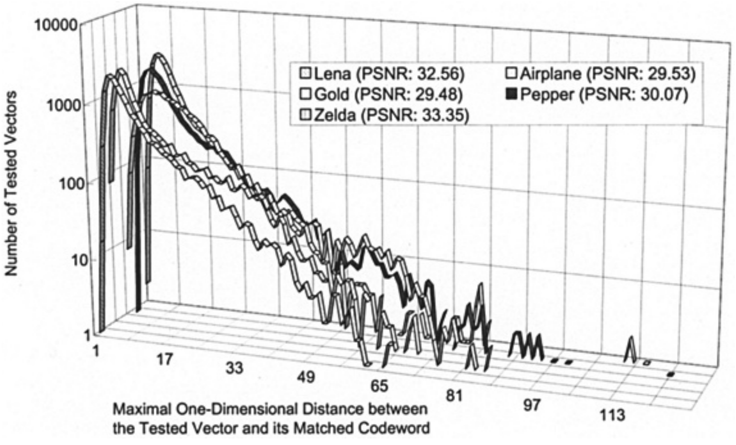


Fig. 1. Distribution of the Maximal One-dimensional Distances for Different Images.

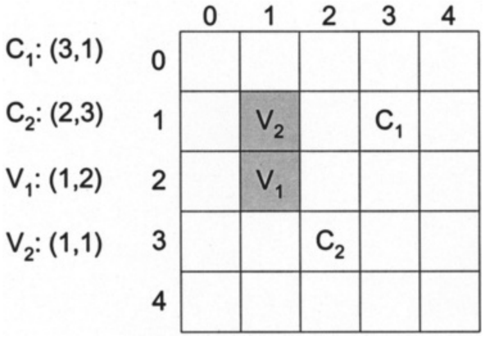


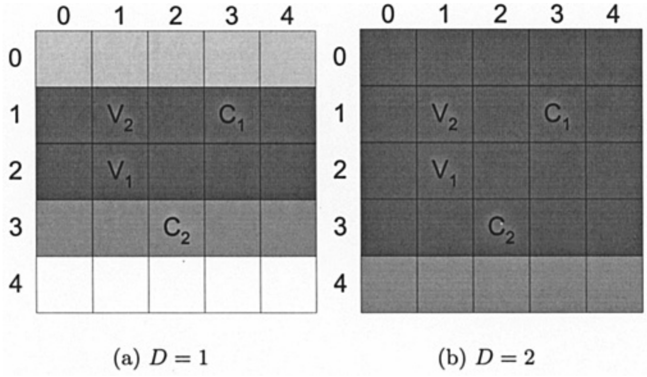
Fig. 2. A Two-dimensional Example.

calculated to decide the one with the smallest Euclidean distance. To nail down the problem, a wider range could be adopted. However, the conjunct bricks are also increased due to the larger squares, as shown in Fig. 3(b).

A suitable range is thus critical to the performance of the proposed scheme since a wider range could increase the number of candidates while a narrow range might cause null set. In our experiments, various ranges are investigated to evaluate the performance and the image quality. Consequently, the construction/lookup procedure of the searchable data structure "THRESHOLD" is introduced.

2.1 The Construction of the *THRESHOLD*

The proposed data structure *THRESHOLD*, T, contains m entries, which consists two fields T^S and T^E. Assuming dimension j is selected to construct the data structure and D is the pre-defined range. T_p^S indicates the index of the



(a) $D = 1$ (b) $D = 2$

Fig. 3. The Concept of the *THRESHOLD*.

first codeword whose j_{th}-dimensional position is within the range from $p + D$ to $|p - D|$, and T_p^E is the last one, where $0 \leq p \leq m - 1$. Before constructing the data structure, the codewords should be sorted according to the j_{th}-dimensional value to ensure that the codewords within T_p^S and T_p^E could conform the range definition. The required storage for *THRESHOLD* is $2m\log_2 N$ bits. For a 256-gray-level vector-quantization encoder with 256 codewords, the required storage is 512 bytes.

```
The Construction Algorithm
INPUT: The sorted codewords
OUTPUT: The Constructed THRESHOLD
For each position $p$ BEGIN
  $T_p^S$ =the smallest index $i$ of the codewords, which satisfy $|p - D| \leq c_{i(j)} \leq p + D$.
  $T_p^E$ =the largest index $i$ of the codewords, which satisfy $|p - D| \leq c_{i(j)} \leq p + D$.
END
```

The lookup procedure combines fast pruning by *THRESHOLD*. Assuming the TLUT is combined to fasten the calculation. Only the codewords whose index x satisfies $T^S \leq x \leq T^E$ are calculated in vector quantization. For the quantized vector x, the x_jth entry of the *THRESHOLD*, $T_{x_j}^S$ and $T_{x_j}^E$, is fetched. Each codeword whose index i satisfies $T_{x_j}^S \leq i \leq T_{x_j}^E$ will be selected as the candidate, and the Euclidean distance is calculated by coupling TLUT. The pseudo code for lookup procedure is listed below.

For each vector quantization, only two extra memory accesses are required. Since there is no extra computation cost, adopting *THRESHOLD* is simple and suitable for hardware implementation. It can also combine with state-of-the-art VQ algorithms to further improve the performance.

```
The Proposed VQ Algorithm
For each vector $x$ BEGIN
  Fetch the $T^{S}_{x_j}$ and $T^{E}_{x_j}$.
  For each codeword $i$, where $T^{S}_{x_j} \leq i \leq T^{E}_{x_j}$
  BEGIN
      Calculate Euclidean distance $d(x, c_i)$
      $d(x, c_i) = \sum_{j=0}^{k-1} TLUT[|x_{(j)}, c_{i(j)}|]$.
      If $d(x, c_i) \leq min\_distance$ BEGIN
         $min\_distance\_id = i$
         $min\_distance = d(x, c_i)$
      END
  END
  $min\_distance\_id$ is the result for $x$.
END
```

3 Performance Evaluation

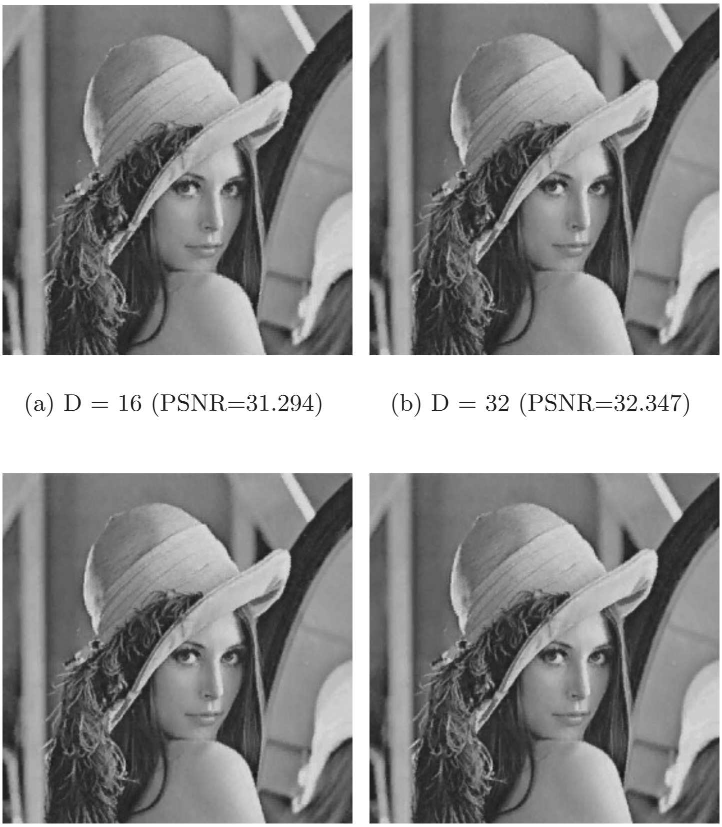
We have conducted several simulations to show the efficiency of the proposed scheme. All images used in these experiments were 512 × 512 monochrome still images, with each pixel of these images containing 256 gray levels. These images were then divided into 4×4 pixel blocks. Each block was a 16-dimensional vector. We used image "Lena" as our training set and applied the Lindo-Buzo-Gray (LBG) algorithm to generate our codebook C. In the previous literature [1,2], the quality of an image compression method was usually estimated by the following five criteria: compression ratio, image quality, execution time, extra memory size, and the number of mathematical operations. All of our experimental images had the same compression ratio. Thus only the latter four criteria are employed to evaluate the performance. The quality of the images are estimated by the *peak signal-to-noise ratio* (PSNR). A larger PSNR value indicates better preserved the original image quality. The extra memory denotes the storage needed to record the proposed scheme and TLUT. As for the mathematical operations, the number of calculated codewords is considered since the operations for each codeword are identical.

The experiments were performed on an IBM PC with a 500-MHz Pentium CPU. Table 1 shows the experimental results of the proposed scheme. VQ indicates the vector quantization without any speedup. The ranges for the *THRESHOLD* vary from 16 to 128. With a smaller range, the image quality is degraded since the occurrence of false matches is increased. Nevertheless, the calculated codewords are reduced by the *THRESHOLD*, the execution time is lessened as well. VQ requires no extra storage while the TLUT needs 256 bytes. The extra storage is 512 bytes for the *THRESHOLD* and 256 bytes for TLUT. The decompressed images are shown in Fig. 4. While the *THRESHOLD* range is enlarged to 64, the image quality is almost identical to VQ and TLUT.

Table 2 illustrates the performance of the proposed scheme based on different images. For the images with better compression quality in full search, the proposed scheme generates more candidates since the codewords are usually close to

Table 1. The Performance of the *THRESHOLD*.

			The Proposed Scheme				
Lena	VQ	TLUT	D=16	D=32	D=48	D=64	D=128
PSNR	32.563	32.563	31.294	32.347	32.531	32.554	32.563
Execution Time (sec.)	1.302	1.091	0.201	0.371	0.52	0.671	1.011
Calculated Codewords	256	256	48	92	130	164	243
Memory Size (byte)	0	256	512 (*THRESHOLD*) + 256 (TLUT)				



(a) D = 16 (PSNR=31.294)

(b) D = 32 (PSNR=32.347)

(c) D = 64 (PSNR=32.554)

(d) D = 128 (PSNR=32.560)

Fig. 4. The Decompressed Lena Images of the Proposed Scheme.

the compressed blocks. While the range is enlarged to 32, the proposed scheme could derive compressed images with comparable quality to full search while requiring only fourth time.

Table 2. The Performance of the *THRESHOLD* based on Different Images.

	Lena			Airplane			Gold			Pepper			Zelda		
FS	256	1.30	32.56	256	1.30	29.53	256	1.30	29.48	256	1.29	30.07	256	1.30	33.35
TLUT	256	1.09	32.56	256	1.11	29.53	256	1.10	29.48	256	1.10	30.07	256	1.09	33.35
D=16	48	0.20	31.29	35	0.14	28.56	44	0.17	29.08	43	0.16	29.13	50	0.21	32.95
D=32	92	0.37	32.35	65	0.26	29.28	87	0.32	29.45	83	0.33	29.66	96	0.36	33.27
D=48	164	0.67	32.55	133	0.51	29.53	160	0.63	29.48	159	0.62	30.04	171	0.66	33.34
D=128	243	1.01	32.56	224	0.90	29.53	240	0.96	29.48	238	0.96	30.07	246	1.01	33.35

4 Conclusion

In this study, we propose a new data structure, *THRESHOLD*, for fast codebook search. The proposed scheme adopts simple data structure with merely two integers to represent the geometrical information. By setting a given range, the proposed scheme could sift out codewords easily, and the proposed scheme is suitable for hardware implementation. The extra storage is 512 bytes and the performance could be improved with a factor of four. In the future, we will apply this concept to the codebook training.

References

1. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston, MA: Kluwer, 1992.
2. T. S. Chen and C. C. Chang, “An Efficient Computation of Euclidean Distances Using Approximated Look-Up Table,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 594–599, June 2000
3. G. A Davidson, P. R. Cappello and A. Gersho, “Systolic architectures for vector quantization,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1651–1664, Oct. 1988
4. H. Park and V. K. Prasana, “Modular VLSI architectures for real-time full-search-based vector quantization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, pp. 309–317, Aug. 1993
5. P. A. Ramamoorthy, B. Potu and T. Tran, “Bit-serial VLSI implementation nof vector quantizer for real-time image coding ,” *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1281–1290, Oct. 1989
6. S. A. Rizvi and N. M. Nasrabadi, “An efficient euclidean distance computation for quantization using a truncated look-up table,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 370–371, Aug. 1995

An Efficient Object Based Personal Video Coding System

Cataldo Guaragnella[1] and Tiziana D' Orazio[2]

[1] Politecnico di Bari, DEE - Electrics and Electronics Department
4, Via E. Orabona, 70125 Bari - Italy
guaragnella@poliba.it

[2] Institute of Intelligent Systems for Automation - C.N.R.
Via Amendola 122/D-I, 70126 Bari, Italy
dorazio@ba.issia.cnr.it

Abstract. A motion based unsupervised neural network approach to motion segmentation is addressed, and embedded in an automatic object based coding system. The motion estimation phase is carried out by an arbitrarily shaped object oriented block based technique (S-BMA). An efficient polynomial motion model is used to describe motion fields and jointly segment images into background-foreground. The proposed technique is embedded in a H.263-like coding system and tested on the foreman sequence. Preliminary results on standard video sequence seem promising.

1 Introduction

Several coding applications need simple and effective motion segmentation procedures: real time MPEG-4 based coding structures need automatic segmentation procedures, to split an image into objects moving in the scene.

Personal video communication systems need simple procedures to segment foreground objects from the background. All the segmented objects are coded separately and multiplexed together in the whole stream.

The simplicity and good performances of Block Matching (BM) based motion estimation algorithms have favored their use in most coding standards as MPEG 1-4 and H.261-3 ([1-7]). BM algorithms allow low power prediction error even if wrong optical flow estimates occur: i.e., for outdoor video sequences illumination variation may occur due to the exposure variation of the main object in the scene to the dominant light source.

In this work an automatic Background-Foreground segmentation technique is presented: a block based optical flow field estimation taking care of possible illumination variation fields ([15]) is used. Motion estimates, together with color and other side information are arranged into pixel based vector information; an unsupervised neural network ([14]) is used to split the whole image into two clusters. Due to coherence of color information and motion, fine clustering of the moving object is obtained. The procedure has been embedded into a H.263-like coding system, proposed for wireless personal video coding applications.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 655–664, 2004.

2 S-BM and Illumination Variation

The improvement in prediction error, easily achievable using S-BM ([15]) can be further enhanced by taking care of the possible presence of illumination variations inside the video sequence.

To face this problem without reducing the algorithm efficiency too severely, a multiplicative illumination variation model can be conveniently chosen: in this case the estimation procedure requires only the evaluation of the illumination coefficient at each trial motion vector.

Here a linear illumination variation model is used, and the motion is estimated using the S-BMA described in [16]. Each block whose motion has to be estimated is initially split into at most three areas, obtained by the thresholding the frame difference; three areas, P, N, Z, are hence defined as:

$$\begin{cases} P = \{(x,y) | FD(x,y) > Th, \quad 0 \leq x \leq H-1, 0 \leq y \leq V-1\} \\ N = \{(x,y) | FD(x,y) < -Th, 0 \leq x \leq H-1, 0 \leq y \leq V-1\} \\ Z = \{(x,y) | \, |FD(x,y)| < Th, \ 0 \leq x \leq H-1, 0 \leq y \leq V-1\} \end{cases} \tag{1}$$

Here Th is a defined threshold level. To take care of illumination variation due to different exposure of the moving object in a dominant light source, the typical situation in personal video communication systems, the prediction error function must be defined differently.

Let us define a cost function, for each domain D inside each block as:

$$E\left(v_x, v_y, \alpha\right) = \frac{1}{\|D\|} \sum_{(x,y)\epsilon D} |I_t(x,y) - I_{t-1}(x+v_x, y+v_y) \cdot \alpha| \tag{2}$$

Here α is the illumination variation coefficient to be estimated on the sequence for each sub-block region whose motion is searched, and $\|D\|$ represents the number of pixels inside the sub-block detected coherently moving area (P,N or Z). The multiplicative illumination variation model allows to estimate α at each trial motion vector as the scale coefficient able to minimize the mean squared prediction error. With the introduced symbols, we choose α so that the modified MSE (m-MSE), defined as:

$$E^2\left(v_x, v_y, \alpha(v_x, v_y)\right) = \frac{1}{\|D\|} \sum_{(x,y)\epsilon D} |I_t(x,y) - I_{t-1}(x+v_x, y+v_y) \cdot \alpha(v_x, v_y)|^2 \tag{3}$$

reaches the least value for the trial motion vector. The implemented technique, at each trial motion vector, computes the α value as:

$$\alpha(v_x, v_y) = \frac{\sum_{(x,y)\epsilon D} |I_t(x,y) \cdot I_{t-1}(x+v_x, y+v_y)|}{\sum_{(x,y)\epsilon D} |I_{t-1}(x+v_x, y+v_y)|^2} \tag{4}$$

which is obtained as a result of the least squares approach with respect to the illumination parameter on the coherent moving area inside each sub-block P, N or Z areas.

The least m-MSE value, determined on the whole set of possible displacements in the search domain, defines jointly the motion vector and the illumination variation coefficient estimates.

3 The Implemented Coding Structure

In this paragraph we describe the coding structure oriented to personal video communication applications, where a principal moving video object is assumed acting on a fixed or slowly moving background.

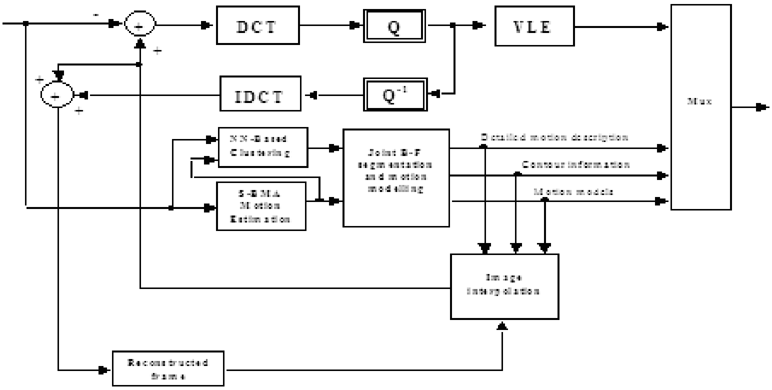


Fig. 1. The proposed H.263-like coding system

Several techniques for video objects segmentation to be used in coding applications have been presented in literature; all such segmentation techniques are based on the soft-thresholding of the motion fields [8-13]. Joint techniques to obtain both motion modeling and image segmentation into objects, often result computationally expensive.

Video objects segmentation is not an easy task due to inaccuracy of motion knowledge specially on moving object border blocks. The segmentation obtained exploiting only motion information is inaccurate, but it can be enhanced by the joint use of several information at hand, like color and motion.

Instead of heuristic considerations, a locally connected unsupervised neural network (NUSD, [14]) is adopted to simplify the information at hand, in order to obtain the image partitioned into cluster maximally coherent in terms of their information content. After the image partition, each detected cluster is assigned to one between two possible image sub-sets corresponding to 'foreground' and 'background' hypothesis. This second step consists of a merge procedure: polynomials have been used to describe the motion models for each area the image should be split into (background and foreground). The estimation of the polynomials coefficients is performed jointly with the definition of the image segmentation.

The procedure uses iterated trials: at the generic trial, a given cluster is chosen to be joined either to the background or to the foreground. Cluster aggregation is based on the "closeness" of the motion model of background or foreground to the motion model of the candidate cluster: the least prediction error obtained interpolating the candidate cluster by the two motion models of background and foreground defines its inclusion. Once the inclusion has been accepted, a new motion model is computed for the newly defined image subset.

3.1 The Proposed Object-Based Coding Structure

The block diagram of figure (1) represents the H.263-like coding system modified to accommodate the proposed automatic system for motion segmentation and object based coding. The automatic system for background–foreground segmentation used as an useful coding application uses a neural network and a merge procedure to obtain the final image partition into background and foreground. The clustering algorithm used to simplify the image is synthetically described hereafter, while in a successive sub-paragraph a brief description of the merge procedure to obtain the final segmentation is given.

Once obtained the motion estimates based on S-BM algorithm shortly described in preceding section, the motion estimates, together with color information are arranged in vectors and used in the unsupervised NN clustering procedure to obtain compact clusters locally coherent in color and motion information. The obtained clusters are now treated as single entities and arranged in a partition of the image whose goal is the segmentation of the image into foreground (moving) object and background.

The proposed coding structure is a complete H.263-like coder, with exception for the motion information coding system and the interpolating system.

3.2 Locally Connected Unsupervised NN

The used NN is hereafter shortly described; a complete description can be found in [14]. It performs several successive densities measures in a hyper-sphere data observation window, computes the centroid of the local "mass distribution" of the input data and updates the centroid location at the local detected centroid of the observed data. This way the concentration gradient is pursued until the network reaches a convergence in a local steady-state point of the data structure. At each step the data observation window is reduced in accordance with the increased data density to allow a higher resolution in the detection of the data feature.

A schematic block diagram of a single unit of the used neural network is reported in the left part of figure 2.

Once the steady-state point has been reached, a new neuron can be initialized. This time it will pursue a new data feature basing on two force-sources: an attracting gravitational force, due to the higher concentration detected in a given hyper-sphere window, as in the first neuron case, and a repulsive one, due

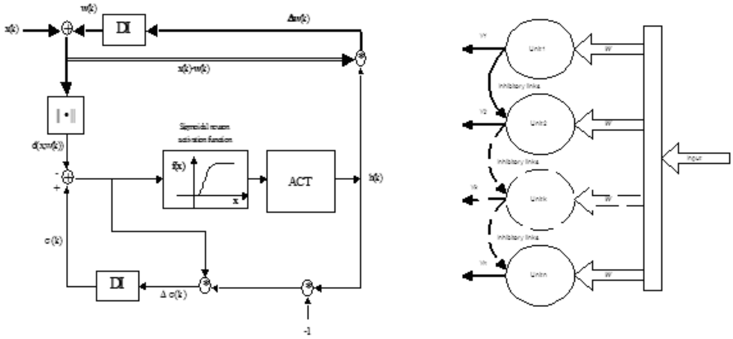


Fig. 2. Left: NUSD block diagram; Right: N-NUSD scheme of application

to the already acquired neuron. The two forces balance in order to allow a new feature detection.

The clustering algorithms used to segment the data on the basis of the acquired data features is obtained by splitting the whole data structure by means of geometrical considerations.

3.3 Background-Foreground Video Segmentation and Polynomial Modelling of Motion Fields

The technique used to obtain the desired video Background-Foreground (BF) segmentation was presented in [16]; it uses two different motion models for the background and the foreground moving object: 2^{nd} order polynomial motion have been used here. Such models are able to well describe perspective point of view changes of rigid objects, but also represent a good choice to model complicated motion fields.

The motion model expression is reported in (5) for the background models; similar equations are valid for the Foreground. a_i and b_i represent the model coefficients for respectively the horizontal and vertical motion fields (respectively, D_h and D_v). Each cluster detected in the first phase of image simplification is segmented into connected regions, and relabelled as different image clusters.

$$\begin{cases} D_h(x,y,t) = a_{B,1}(t) + a_{B,2}(t) \cdot x + a_{B,3}(t) \cdot y + \\ \qquad +a_{B,4}(t) \cdot x^2 + a_{B,5}(t) \cdot y^2 + a_{B,6}(t) \cdot x \cdot y \\ D_v(x,y,t) = b_{B,1}(t) + b_{B,2}(t) \cdot x + b_{B,3}(t) \cdot y + \\ \qquad +b_{B,4}(t) \cdot x^2 + b_{B,5}(t) \cdot y^2 + b_{B,6}(t) \cdot x \cdot y \end{cases} \tag{5}$$

The problem of BF video segmentation can be reduced to a redistribution of the clusters obtained by the neural network into two disjoint sets. The algorithm evolves selecting the two starting clusters, C_B (background) and C_F (foreground) initially chosen as the two maximally different in terms of their average motion. For all the pixels inside the cluster a polynomial motion model is computed in the mean square error sense. For a generic cluster C, the aggregation step consists in computing the membership function values for the two motion models at hand. The membership function has been defined basing on the prediction error computed on the whole cluster, C, by the application of the two computed motion models. The least prediction error decides the cluster inclusion inside the growing (background or foreground) region. Once the cluster is accepted into a region, a new motion model is recomputed, and the procedure is iterated until the whole image clusters are aggregated to either background or foreground.

At the end, motion models, together with the image segmentation information, represent the side information to be used in the first layer of the motion description system in the used coding structure. The estimated polynomial models cannot take care of detailed motion in presence of real scenes for eyes or mouth. To reduce the DFD of the prediction obtained using motion models, for a few blocks in the image, a detailed motion description is also used. It is based on the motion and illumination variation description of the sub-block areas of the image portions not properly described by the estimated motion models.

The information required to code this detailed motion consists an index for the block position in the image together with the motion vector and (eventually) the illumination variation coefficient. Coding requirements are very modest if compared to standard MPEG coding technique, as only information about the foreground object contour is required and few coefficient to describe the motion models.

As expected the good appearance of the reconstructed image, though still not loss-less coded, presents no artifact due to the block based motion field estimation and a fine foreground object segmentation.

3.4 Coding of Motion Information: Polynomial and Detailed Motion Description

The bit rate required to describe the whole image with the proposed coding scheme needs the information described hereafter:

- the description of both background and foreground motion models: $2 \times (2 \times 6)$ coefficients to describe the quadratic polynomial motion field;
- the description of the contour of the foreground object;
- the detailed motion blocks required for coding (position, shape and motion vector(s));
- the coded displaced frame difference.

The main object contour description requires the centroid position on the image and a sub-sampled contour description, arranged in a chain coding system (see fig. 3). Position and motion vectors for detailed motion description of a few blocks on the image (less than 10%).

The description of each component of the motion vector requires $log_2(2 \cdot Max\_Motion/Step)$ bits, where Max_Motion is the unilateral block matching search domain, and $Step$ is the used step in the search procedure.

The displaced frame difference coding requires the same amount of bits for the standard MPEG coding structure as the statistics of the prediction error is roughly the same.

In MPEG based bit streams transmission errors occurring in the motion description produce bad artifacts on the decoded image, particularly when the motion field is coded in a differential form, because an error in the motion vector can propagate across the whole image and produce effects over several frames. In this approach, instead, due to models, errors produce deformations in predicted image, but no blocking effects.

The residual error bit rate can be coded efficiently using the FGS wavelet/DCT decomposition in order to make the bit stream adaptive to the channel congestion situations.

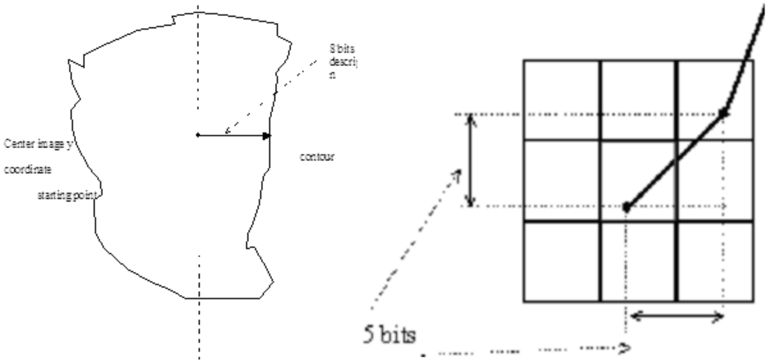


Fig. 3. visualization of the chain coding implemented technique for the foreground object: chain coding is implemented connecting points of the line separating Background and foreground regions with segments whose ending points relative position are described by 5 bits

The main difference between the proposed coding system and the standard MPEG lies in the unavoidability of DFD transmission for block based perceptual coding when applied in very low bit rate applications: indeed low bit rate require few bits for the residual error coding, thus blocking artifacts appear on the decoded image, worsening very rapidly the reconstructed video quality with the bit rate decrease.

The predicted images, obtained with polynomial models, produce good quality images, even without any DFD transmission, so that the proposed coding system candidates as a possible solution to the problem of quick image degradations in rate adaptive video coders for highly variable channel capacity systems; even neglecting the DFD information, the reconstructed video reveals clearer in the image content, but less fluid in motion evolution representation.

4 Experimental Results

Tests have been conducted on the standard "foreman" sequence. Here illumination changes of various regions of the face of the actors moving on the background are real, and mainly due to motion. All results refer to the sequence in Quarter Intermediate Format (YUV-QCIF, 144 × 176 pixels), with a resolution of 8 bits per pixel/color. The motion estimation has been computed using full search algorithm with step size of 0.5 pixels. Bilinear interpolation has been used to compute predictions at half pixel accuracy. A fixed threshold value ($Th = 4$) has been used for all frames of each processed sequence. In the proposed application the illumination variation information was neglected (video semantic is preserved with motion compensation only). Illumination variation has been added as an extra information for the detailed motion description, for less than 10% of the total number of image blocks.

Figure 4 shows results of the proposed algorithm: the top-left image is the foreground object prediction, the top-right one is the motion compensated frame difference obtained using only polynomial motion models (without detailed motion description). The Bottom-left and right images of figures 4 represent the motion models (amplified for visualization). As it can be noted, the motion compensated frame difference presents a smooth appearance, evidencing the total absence of blocking artifacts; the highest peaks of the motion compensated frame differences reveal a not exact motion field description, in some parts of the image; this problem is mainly due to the large area characterized by the same motion field. Better results would require higher order polynomials or smaller regions. In both cases if prediction is enhanced, the motion description bit rate would increase.

The impossibility of separating the hat of the foreground speaker from the background is an evidence of this case: the problem resides in the closeness of the hat color with the background, and in the smoothness of the hat. The color closeness make indistinguishable the foreground area from the background in the color space, so that only motion can help in separating such features. Unfortunately, the object smoothness reduces very much the resolution capability of a block based motion technique, so that both motion estimates and color components fall in the background subspace, and the neural network approach fails in this case.

Better performances might be obtained in a multi-resolution approach to motion estimation.

5 Discussion and Conclusions

Presented results seem to indicate that the "Slicing" modification of the block matching is a suitable mean for obtaining better motion estimates than other block based techniques to be used in motion segmentation applications: the slicing of a single block, multiplies the number of motion vectors to describe the whole prediction, but requires an efficient method to describe the segmented

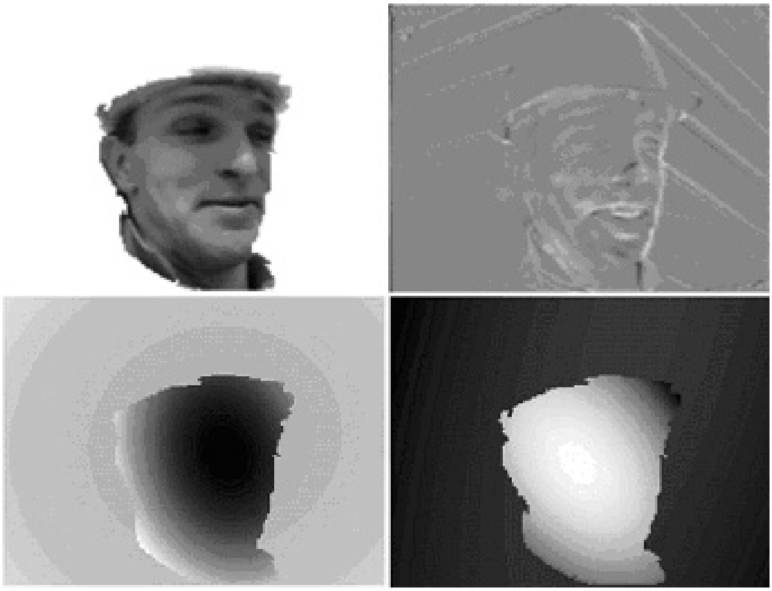


Fig. 4. Top left: The obtained image segmentation; Top-right: displaced frame difference (amplified for visualization); Bottom-left: obtained polynomial Motion Field (horizontal) Bottom-right: (vertical)

areas inside each block. This BM algorithm reveals well suited for motion segmentation applications such as in object based video coding where an efficient technique for motion estimation makes the motion segmentation reliable even in real time coding environments as in personal video communication systems.

In internet applications, severe network congestion situations, often cause freezes in the received video. Freezes probability can be much lowered, dropping residual coding information with acceptable received image quality. The side motion information to obtain the predicted image is minimal and higher performances with respect to standard techniques appear evident mainly with small block dimensions, even in rate adaptive FGS coding systems.

The main drawback of the proposed coding structure resides in the lower fluidity of the motion of the coded sequence at low bit rates.

References

1. ISO/IEC JTC1 CD11172, Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 Mbit/s–Part 2: Coding of Moving Picture Information, ISO, 1991
2. ISO/IEC DIS 13818-2, Information Technology - Generic Coding of Moving Pictures and Associated Audio Information - Part 2: Video, ISO, 1994

3. T. Ebrahimi, MPEG-4 Video Verification Model: a video encoding/decoding algorithm based on content representation, Signal Processing: Image Communication, Dec. 1998
4. F. Pereira, MPEG-4: a New Challenge for the Representation of Audio-Visual Information, Proc. of Picture Coding Symp., Melbourne, 1996
5. CCITT SG XV, Recommendation H.261 - Video Codec for Audiovisual Services at px64 kbit/s, COM XV-R37-E, Int. Telecommunication Union, August 1990
6. ITU–T Draft, Recommendation H.263 - Video Coding for low bit rate communication, Int. Telecommunication Union, November 1995
7. L. Chiariglione, The development of an integrated audiovisual coding standard: MPEG, Proceedings of IEEE, Vol. 83, No. 2, February 1995
8. N. Diehl, Object oriented motion estimation and segmentation in video sequences, Signal Processing: Image communication, vol. 3, pp. 23–56, 1991
9. Huang, Y., Zhuang, X.H., Yang, C.S., Two Block-Based Motion Compensation Methods for Video Coding, IEEE Trans. on Circuits and Systems for Video Technology vol. 6, No. 1, pp. 123–126, 1996
10. P. Eisert and B. Girod, Illumination Compensated Motion Estimation for Analysis Synthesis Coding, Proc. of 3D Image Analysis and Synthesis, pp. 61–66, Erlangen, November 1996
11. P. Eisert and B. Girod, Model-based Coding of Facial Image Sequences at Varying Illumination Conditions, 10th IMDSP Workshop 98, pp. 119–122, Alpbach, 1998
12. P. Salembier, F. Marques, 1999, Region based representation of image and video: Segmentation tool for multimedia services, IEEE Trans. Circuits and systems for Video Technology, invited paper, vol. 9 no. 8, Dec. 1999
13. P.R. Giaccone, D. Tsaptsinos and G.A. Jones, 2000, Foreground-Background Segmentation by Cellular Neural Networks, Proceedings of the International Conference on Pattern Recognition (ICPR'00)
14. G. Acciani, E. Chiarantoni, and M. Minenna, A new non Competitive Unsupervised Neural Network for Clustering, Proc. of Intern. Symp. On Circuits and Systems, Vol. 6, pp. 273–276, London May 1994
15. C. Guaragnella, E. Di Sciascio. Object Oriented Motion Estimation by Sliced-Block Matching Algorithm, IEEE Proc. of ICPR '2000, Intl. Conference on Pattern Recognition, Barcelona, 2000
16. G. Acciani, E. Chiarantoni, C. Guaragnella, V. Santacesaria, 2002, Neuro-Fuzzy Architecture for Background-Foreground Video Segmentation, Proc. NF 2002, Cuba, 6-10 Jan. 2002

Multiview Video Coding Based on Global Motion Model

Xun Guo[1] and Qingming Huang[2,3]

[1] Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[3] 3Graduate School of Chinese Academy of Sciences, Beijing, China
{xguo,qmhuang}@jdl.ac.cn

Abstract. In this paper, we present a novel scheme for coding multiview video sequence based on global motion prediction between adjacent views. For that, the left-most view is compressed as reference sequence using standard block-based motion compensated prediction coding. And its right view is compressed with global motion prediction from the left view images. In the prediction, an eight-parameter global motion model is used to compute the global motion information between left and right-view images. Then the motion vectors of the right-view image are predicted from the left-view image based on the global information. To further reduce the coding complexity and improve coding efficiency, macroblock modes of current image are also predicted from the left-view image. H.264 coding scheme is employed as the baseline, in which Rate-Distortion Optimization is used to select the best coding mode. Experimental results show that, compared to coding multiview video sequence independently, the proposed scheme can save the bitrate up to 15%.

1 Introduction

Multiview video is one of the main categories of 3D video, which can give users the opportunity to choose their favorite viewpoints freely. Multiview video consists of video sequences captured by N cameras at the same time but different positions. Because it contains multiple sequences, it requires much more bandwidth than traditional 2D video in transmission. Therefore, how to compress multiview video sequence efficiently has become popular due to its wide application.

In the past years, some work has been done on multiview video coding. In [1]-[2], Grammalidis uses a multiview matching cost as well as pure geometrical constraints algorithm to estimate disparity and to identify the occluded areas in the views. And in [3], a sprite generation algorithm in multiview sequence is proposed to improve coding efficiency. But all these methods are based on the sequences for videoconferencing. Up to now, many video coding standards have been established such as MPEG-2, MPEG-4, H.263 and H.264 [4]. However,

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 665–672, 2004.

none of them really supports multiview video coding. Recently, MPEG/3DAV group has started the work on 3dav standard. More and more people begin to concern multiview video coding. Some preliminary study and experimental results of multiview video coding have been reported. There are mainly three coding methods are investigated [5]-[7]. A straightforward idea is to code the N multiple video sequences separately. In this method, only temporal correlation within one sequence is used. Another method is to utilize inter-view correlation only. In this case, images of one view are only predicted by their left view images. The third method utilizes both temporal and inter-view correlation. Test results show that the third method is not so efficient when only simple block-based motion compensated prediction is used to exploit inter-view correlation. Due to the high similarity and the little displacement between two adjacent views, global motion information can be used to improve the coding efficiency.

In this paper, we propose an efficient multiview video coding scheme based on global information between two adjacent views. JVT coding scheme is employed as the baseline. In order to find out the inter-view correlation, two adjacent views are coded as reference and secondary sequences respectively. To fully utilizing the correlation between the two views, we propose to predict the motion vectors of the right-view images from that of the left-view images based on the global motion information between them. And to further reduce the computational complexity in mode decision, we also propose to predict the current macroblock mode in right-view image from that of the left-view image.

The remainder of this paper is organized as follows. Section 2 describes the multiview video coding scheme in detail. Section 3 describes the key technologys of the proposed scheme including motion vector prediction based on global motion estimation and macroblock mode prediction method between left-view and right-view. In section 4, some experimental results of proposed scheme are given and conclusions are drawn in section 5.

2 Multiview Video Coding Scheme

As we know, there are two kinds of motion in most video sequences: local motion and global motion. In MPEG-4 coding scheme [8], global motion compensation (GMC) technology is used to indicate global motion with a set of parameters. In H.264, great achievements have been reached by dividing the traditional 16x16 macroblock into smaller blocks. Therefore, GMC is less competitive than local motion compensation (LMC) even in the case global motion existing due to inaccurate global motion estimation (GME). However, multiview video coding is different from monoview video coding. Due to the high correlation and similarity of adjacent view images, the displacement between left-view and right-view is quite similar to the global motion between two images in one sequence. This characteristic makes the motion prediction between the corresponding images in the two sequences possible.

Fig.1 gives the proposed multiview video coding structure. Left-most view and its right view are coded as reference sequence and secondary sequence respectively. As shown in Fig.1, the two video streams are coded using GOP structure. P and B frames in secondary sequence are coded by referencing previous reconstructed frames in its own sequence or the corresponding frame in the reference sequence.

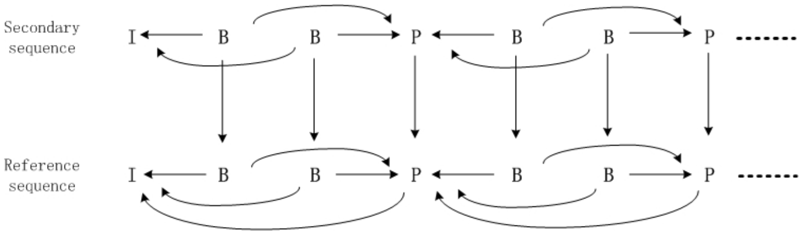


Fig. 1. The proposed multiview video coding structure

The goal of the inter-view prediction is to get the global information between the right-view image and the left-view image. Traditional eight-parameter global motion model (perspective motion model) is used in our scheme.

Fig.2 shows the proposed multiview coding scheme based on JVT. Reference and secondary sequences are coded using standard JVT coding structure. GME, MV prediction and macroblock mode prediction are added to this framework.

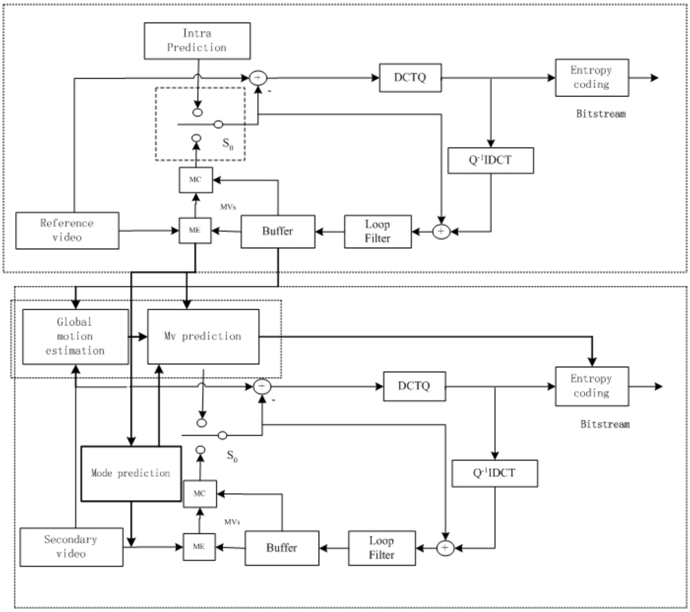


Fig. 2. The proposed multiview video coding scheme based on JVT

When the secondary sequence is coded, GME between the image to be coded and the corresponding left-view image is performed first using perspective motion model. Then the macroblock mode is predicted from the macroblock in the same position in corresponding left-view image. Now, motion vector prediction (MVP) is performed based on the global motion information and the MV of the corresponding left-view macroblock. For each macroblock, the prediction can be obtained by LMC and MVP mode. Rate-Distortion optimization algorithm is used to select the best mode.

3 Key Technologies of the Multiview Video Coding

Motion vector prediction and macroblock mode prediction are the key technologies of the proposed multiview video coding scheme. Motion vector prediction can utilize the high correlation between adjacent views efficiently. And macroblock mode prediction can reduce the coding complexity of LMC.

3.1 Motion Vector Prediction

In H.264 inter frame coding, there are 7 block sizes, 16x16, 16x8, 8x16, 8x8, 8x4, 4x8 and 4x4. Thus, for P and B frames, macroblock mode can be selected from SKIP, 16x16, 16x8, 8x16 and 8x8 mode. When the 8x8 mode is chosen, each 8x8 block can be divided into smaller sub-block including 8x4, 4x8, and 4x4 sizes. INTRA-16x16 and INTRA-4x4 modes are also used for inter frame prediction.

If the motion vectors of current encoding frame can be predicted from its corresponding left-view image using GME, large number of bits for coding motion vectors can be saved by only coding the global motion parameters. For this purpose, motion vectors of the secondary images are computed from that of the reference image. This can be accurate because of the little disparity of the two images.

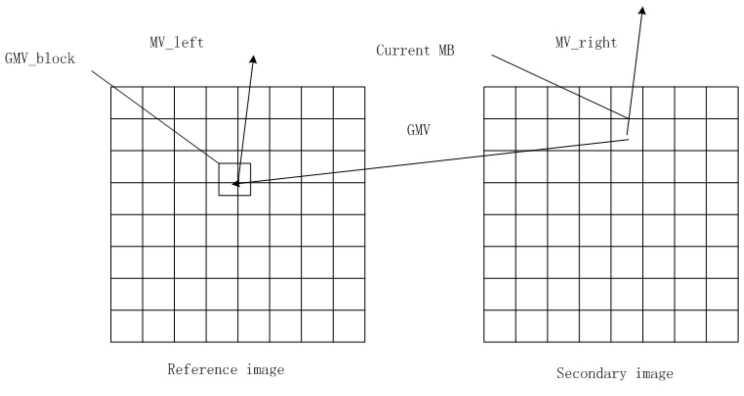


Fig. 3. The process of the MV prediction based on global motion vector (GMV)

Fig.3 shows the proposed MV prediction algorithm. The two grids denote current image to be encoded in secondary sequence and the corresponding image in reference sequence respectively. GMV is the global motion vector of current-MB computed using the global motion information between the two images. In our scheme, 4 GMV sizes are employed: 16x16, 16x8, 8x16, 8x8. The algorithm of computing MV-right is as follows:

1)Get the macroblock mode of Current-MB through LMC.

2)Determine GMV size of Current-MB according to the macroblock mode. If the mode is 16x16, 16x8 or 8x16, GMV of 16x16, 16x8 or 8x16 block size is computed. Otherwise, 4 8x8 GMVs are computed.

3)For all GMVs in Current-MB, find the prediction blocks position (GMV-block) in reference image.

4)Get the motion vectors of the GMV-block as the prediction vectors of Current-MB. Because GMV-block may locate in non-integer MB position, we take vector of most pixels as the GMV-block vector.

3.2 Macroblock Mode Prediction

In JVT coding scheme, mode decision is one of the most important parts. When inter frame coding is performed, the 16x16, 16x8, 8x16 and 8x8 modes have to be full searched to find the best mode, when 8x8 mode is selected, sub-block search for smaller block size has to be performed. If Rate-Distortion Optimization is opened, the large computational complexity will spend a lot of encoding time.

According to statistics, there are more than 70 percent macroblocks that have the same macroblock mode in two adjacent views. Due to this characteristic, mode prediction can be done to right-view images from left-view images. In proposed scheme, we take the macroblock mode of the left-view image as the prediction of that of right-view image. RD cost in the prediction mode is computed. If the cost is less than a threshold T, the mode prediction is thought reliable. Otherwise, the coding process performs as the normal. By using this mode prediction algorithm, computational complexity in RDO is reduced significantly.

4 Experimental Results

In order to verify the performance of the proposed coding scheme, we present the experiment results on some multiview sequences in this section. Our algorithm is implemented based on H.264 reference software JM 7.6 and JM 7.6 scheme without inter-view prediction is taken as the reference. We use the multiview test sequences Crowd, Race provide by KDDI Lab and Aquarium provided by Tanimoto Lab. In each sequence, 600 frames (YUV 4:2:0, 320x240) are coded.

The purpose of this paper is to exploit the correlation between two adjacent views, so only the left-most view and its right view sequences are coded with our scheme. They are also coded independently using JM7.6. In both algorithms, the

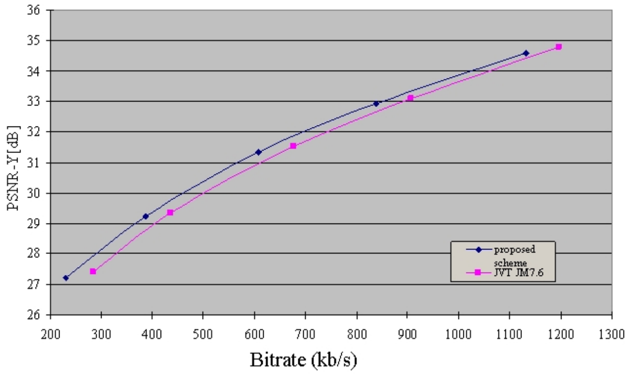


Fig. 4. Coding results of the second view of Crowd

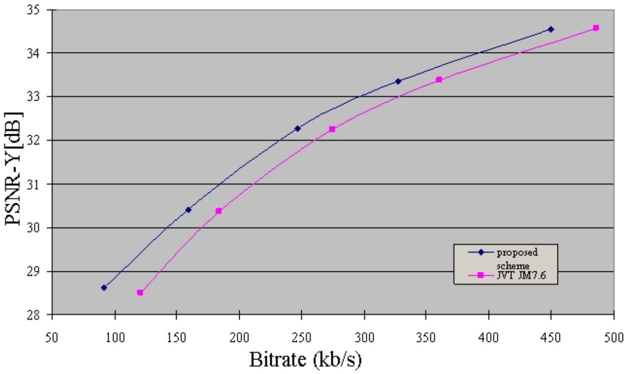


Fig. 5. Coding results of the second view of Race1

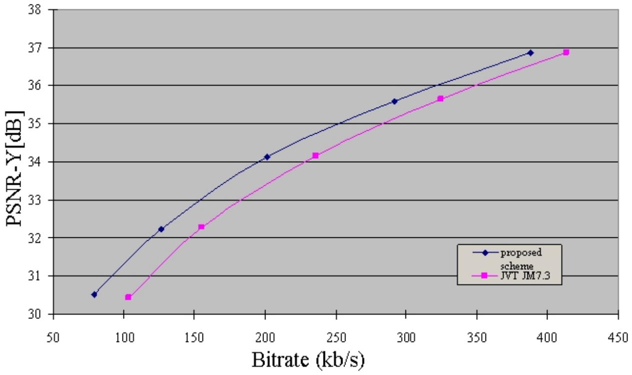


Fig. 6. Coding results of the second view of Aquarium

same parameters, i.e. QP, RDO and reference frame number, are used. Therefore, we can get the coding gain from comparing the two results of the right view.

All these sequences are captured by parallel cameras, which are placed densely (with the distance less than 20 cm). Fig. 4 gives the R-D curves of the second view of Crowd which has high complexity. Fig. 5 and Fig. 6 give the R-D curves of the second view of Race1 and the second view Aquarium which have moderate complexity. Label ”JVT JM7.6” denotes the coding method using JVT JM7.6 reference software. And Label ”proposed scheme” denotes our coding scheme. The figures show that compared to encode the multiview sequences separately with JVT, the proposed scheme has higher PSNR values by about 0.3-0.4 dB for crowd sequence, and 0.8-1 dB at low bit-rate for Race1 and Aquarium. This result show that proposed multiview coding scheme has better efficiency for moderate complex sequences.

5 Conclusions

In this paper, an efficient multiview video coding scheme based on global motion model has been presented. First, an interview prediction coding scheme based on JVT is proposed to encode the left-most view and its right view. Second, the motion vectors prediction algorithm of the right view images are proposed using global motion information between the two view images. Due to the high correlation of the macroblock mode in the two view image, a macroblock mode prediction algorithm is also proposed. Experimental results show that the proposed MV Prediction coding scheme and macroblock prediction algorithm can improve the coding efficiency and the inter-view prediction coding for multiview video is promising.

Acknowledgements. The research of this paper has been supported by National Fundamental Research and Development Program (973) of China (2001CCA03300) and the National Natural Science Foundation of China (No. 60333020).

References

1. N. Grammalidis, M. G. Strintzis: Disparity and Occlusion Estimation in Multiocular Systems and Their Coding for the Communication of Multiview Image Sequences. IEEE Trans. Circuits Syst. Video Technol. Vol.8, pp. 328–344, June 1998
2. N. Grammalidis and M. G. Strintzis: Disparity and Occlusion Estimation for Multiview Image Sequences Using Dynamic Programming. in Proc. Int. Conf. Image Processing (ICIP ’96), Lausanne, Switzerland, Sept. 1996
3. N. Grammalidis, D. Beletsiotis, and M. G. Strintzis: Sprite Generation and Coding in Multiview Image Sequences. IEEE Trans. Circuits Syst. Video Technol., vol.10, pp. 302–311, Mar. 2000
4. 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, no.7, July 2003

5. ISO/IEC JTC1/SC29/WG11: a Multi View Video Compression Scheme Based on Direct View Synthesis. MPEG2003/M10025, Brisbane, October 2003
6. ISO/IEC JTC1/SC29/WG11: Preliminary Study on Multiple View Coding for the Ray Space. MPEG2003/M10054, Brisbane, October 2003
7. ISO/IEC JTC1/SC29/WG11: Ray-Space Coding Using Temporal and Spatial Prediction. MPEG2003/M10178, Brisbane, October 2003
8. ISO/IEC WG11 MPEG Video Group: MPEG-4 video verification model version 16.0. ISO/IEC JTC1/SC29/WG11 MPEG00/N3312, Noordwijkerhout, March, 2000

A Novel Rate-Distortion Model for Leaky Prediction Based FGS Video Coding*

Jianhua Wu and Jianfei Cai

Center for Multimedia and Network Technology,
School of Computer Engineering, Nanyang Technological University,
Block N4, Nanyang Avenue, Singapore 639798
{pg02320384,asjfcai}@ntu.edu.sg

Abstract. In this paper, we propose a novel rate-distortion (R-D) model for leaky prediction based FGS (L-FGS) video coding. The proposed R-D model considers not only the distortion introduced in the current frame but also the propagated distortion due to leaky prediction. The entire system consists of both offline and online processes. During the offline stage, we perform the L-FGS encoding and collect the necessary feature information for the later online R-D estimation. At the online stage, given the transmission bandwidth at that time, we can quickly estimate the R-D curves of a sequence of consecutive video frames based on the proposed R-D model. An excellent property of our proposed R-D model is that even when applying the model for a long video sequence without any update of the actual distortion values, the estimation error is still very small and the error is not accumulated. Experimental results show that the estimated distortion matches the actual distortion very well under different channel conditions.

Keywords: Rate-distortion, leaky prediction, Fine Granularity Scalability.

1 Introduction

Internet video streaming is a very challenging task due to the inherent heterogeneity. For video streaming applications, since different users may access video servers through different access networks, only one bitstream for a video sequence coded at a certain bit rate and stored at the servers cannot satisfy the requirements of different users. Moreover, due to lack of QoS (quality of service) guarantee, essentially, Internet can only provide VBR (variable bit rate) transportation channels for video streaming applications. Therefore, it is desired that video coding itself has the ability to adapt to different users' requirements and the time-varying network conditions. Scalable video coding is one of the common approaches for providing rate adaptation to heterogeneous scenarios.

* This research is partially supported by Singapore A*STAR SERC Grant (032 101 0006).

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 673–680, 2004.

The scalable technique adopted in MPEG-4, i.e., FGS (Fine Granularity Scalability) [1], has received much attention in the past few years mainly because of its simplicity. The basic idea of FGS is to encode a video sequence into two layers: a base layer and an enhancement layer. The base layer is coded without scalability while the enhancement layer is coded bitplane by bitplane. With the bitplane coding method in the enhancement layer, the FGS enhancement layer bitstream can be truncated at any position, which provides fine granularity scalability. FGS is very suitable for adaptive video streaming since it only needs to encode a video sequence once and can adapt to any channel bandwidth within a pre-defined range.

However, FGS is not fit for low bit rate video streaming, e.g., users accessing Internet through bandwidth-limited wireless links. This is mainly because there is no temporal prediction in the FGS enhancement layer, which greatly reduces the coding efficiency, compared with the non-scalable video coding schemes. In order to improve the coding efficiency of FGS, recently, many schemes have been proposed [2,3,4]. The common idea of these schemes is to introduce high quality reference frames to remove the temporal redundancy for the enhancement layer or even the base layer. The robust FGS (RFGS) [4] is the most representative technique, in which, two parameters, the number of bitplanes and the amount of predictive leak, are used to control the construction of reference frames for the tradeoff among efficiency and scalability. In this paper, we denote such leaky prediction [4,5,6] based FGS coding as L-FGS.

Besides introducing better prediction structures, another way to improve FGS coding performance in terms of minimizing overall distortion or having constant video quality is through optimally designing rate control or bit allocation algorithms among video frames, for which the rate-distortion (R-D) relationship for each video frame must be obtained. The problem of obtaining accurate R-D curves for the baseline FGS coding has been solved in [7,8] through either empirical or modelling approaches. In [7], accurate R-D curves are constructed by extracting some R-D points such as the end points of each bitplane during the offline encoding process followed by linear interpolation between neighbor R-D points. In [8], Dai et al. proposed a closed-form R-D model through analyzing the properties of the input to the MPEG-4 FGS enhancement layer, which significantly outperforms the current distortion models such as the quadratic model. However, the problem of obtaining accurate R-D curves for L-FGS is still an open question. The difficulty lies in the dependency among video frames. For example, the R-D curve of the n-th frame depends on the quality of the $(n-1)$-th reference frame, which can only be determined during the streaming stage, i.e., given the bandwidth at that time. To the knowledge of the authors, only approximate approaches have been proposed such as in [9], where a common exponential model is assumed. Those approximate approaches are not accurate and the accumulated errors will become intolerable with the increasing number of the consecutive R-D estimation.

In our previous work [10], we have proposed a model to estimate the distortion propagated from the reference frame due to channel errors for non-scalable video coding. We further applied that model to estimate the propagated distortion due to channel errors for the FGS video coding in [11]. In this paper, we modify the previous model for estimating the propagated distortion due to the varying bandwidth allocated to the reference frame in L-FGS. Moreover, we add in the correlation between the propagated distortion and the distortion introduced in the current frame to the overall R-D model instead of assuming they are independent. This is the major difference from our previous works. The entire system is a combination of offline and online processes. During the offline stage, we perform the L-FGS encoding and collect the necessary feature information for the later online R-D estimation. At the online stage, given the transmission bandwidth at that time, we can quickly estimate the R-D curves of a sequence of consecutive video frames based on the proposed R-D model. An excellent property of our proposed R-D model is that even when applying the model for a long video sequence without any update of the actual distortion values, the estimation error is still very small and the error is not accumulated. Experimental results show that the estimated distortion matches the actual distortion very well under different channel conditions.

2 Overview of Leaky Prediction Based FGS

Using symbol definitions in Table 1, in a typical L-FGS system, the base layer residue $e_B(n,i)$ is obtained by

$$e_B(n,i) = F(n,i) - \hat{F}_B(n-1,j), \tag{1}$$

where $\hat{F}_B(n-1,j)$ is the motion-compensation reference frame. $e_B(n,i)$ and motion vectors are compressed into the base layer. The base layer reconstruction, $\hat{F}_B(n,i)$, is given by

$$\hat{F}_B(n,i) = \hat{F}_B(n-1,j) + \hat{e}_B(n,i), \tag{2}$$

which is stored in the buffer for the encoding of the next frame. After the prediction from the base layer, the enhancement layer data can be represented as

$$\begin{aligned} F_E(n,i) &= F(n,i) - \hat{F}_B(n,i) \\ &= F(n,i) - \hat{F}_B(n-1,j) - \hat{e}_B(n,i). \end{aligned} \tag{3}$$

In the baseline FGS [1], the enhancement layer data is directly compressed into the enhancement layer, i.e,

$$e_E(n,i) = F_E(n,i). \tag{4}$$

However, in the L-FGS, the high quality reference frame is introduced in the enhancement layer with a leaky factor α_n, which can be written as

$$G_E(n,i) = \hat{F}_B(n,i) + \alpha_n \hat{F}_E^p(n,i), \tag{5}$$

where $\alpha_n \in [0,1]$. Note that only the partial data of the reconstructed enhancement layer frame is used as the reference frame, similar to the partial prediction

Table 1. The symbol definitions.

$F(n,i)$	: The original value of pixel i in the n-th video frame.
$\tilde{F}(n,i)$	: The received high quality value of pixel i in the n-th video frame at the decoder.
$\hat{F}_B(n,i)$	: The base layer reconstruction value of pixel i in the n-th video frame in the base layer prediction loop at the encoder.
$G_E(n,i)$	: The high quality reference frame used in the enhancement layer prediction loop.
$e_B(n,i)$	: The base layer residue value of pixel i in the n-th video frame
$\hat{e}_B(n,i)$	: The reconstructed base layer residue.
$F_E(n,i)$	: The enhancement layer data after the prediction from the base layer
$\hat{F}_E(n,i)$	: The reconstructed enhancement layer data in the enhancement layer prediction loop at the encoder.
$\tilde{F}_E(n,i)$	: The reconstructed enhancement layer data at the decoder.
$\hat{F}_E^p(n,i)$	: The partial data of the reconstructed enhancement layer data, which is used in the enhancement layer prediction loop at the encoder.
$\tilde{F}_E^p(n,i)$	: the partial data of the reconstructed enhancement layer data, which will be used in the enhancement layer prediction loop at the decoder.
$e_E(n,i)$	: The enhancement layer residue after all the predictions, including the predictions from both the base layer and the enhancement layer.
$\hat{e}_E(n,i)$	: The reconstructed enhancement layer residue at the encoder.
$\tilde{e}_E(n,i)$	: The enhancement layer residue received at the decoder.

in [4]. This is for a further tradeoff between the coding efficiency and the robustness to drift errors.

The redundancy is further removed by subtracting $\hat{e}_B(n,i)$ from the the difference between $F(n,i)$ and $G_E(n-1,j)$, and the resulted residue is compressed into the enhancement layer. This enhancement layer residue $e_E(n,i)$ can be expressed as

$$e_E(n,i) = F(n,i) - \hat{e}_B(n,i) - (\hat{F}_B(n-1,j) + \alpha_{n-1}\hat{F}_E^p(n-1,j)) \qquad (6)$$

Combining with Eqn. (3) , we simplify $e_E(n,i)$ as

$$e_E(n,i) = F_E(n,i) - \alpha_{n-1}\hat{F}_E^p(n-1,j) \qquad (7)$$

Correspondingly, the enhancement layer reconstruction at the encoder is

$$\hat{F}_E(n,i) = \alpha_{n-1}\hat{F}_E^p(n-1,j) + \hat{e}_E(n,i), \qquad (8)$$

which will be stored in the buffer for the encoding of the next enhancement layer frame at the encoder.

Similarly, as in [4], the high quality reference frame can also be employed in the base layer to further improve the coding efficiency of the base layer. However, in this paper, we only consider introducing the temporal prediction in the enhancement layer, and keep the same encoding scheme in the base layer as the baseline FGS. In addition, we assume that the bandwidth is always enough for transmitting the entire base layer and thus the truncation for rate adaptation only happens in the enhancement layer. H.263+ instead of MPEG-4 is employed to encode the base layer for simplicity. We only consider encoding video sequences with a pattern of one I-frame followed by all P-frames and TMN8 is used as the rate control scheme for the base layer. The L-FGS enhancement layer is encoded bitplane by bitplane, the same as that in the MPEG-4 FGS.

3 Proposed R-D Model

Since we assume the entire base layer of L-FGS can be transmitted without having any distortion, the overall distortion $D(n)$ between $F(n,i)$ and $\tilde{F}(n,i)$ is actually the same as the distortion between $F_E(n,i)$ and $\tilde{F}_E(n,i)$. $\tilde{F}_E(n,i)$ is essentially a random variable and thus $D(n)$ can be expressed as

$$\begin{aligned} D(n) &= E\{[F_E(n,i) - \tilde{F}_E(n,i)]^2\} \\ &= E\{[F_E(n,i) - (\alpha_{n-1}\tilde{F}_E^p(n-1,j) + \tilde{e}_E(n,i))]^2\} \\ &= E\{[(F_E(n,i) - \alpha_{n-1}\hat{F}_E^p(n-1,j) - \tilde{e}_E(n,i)) + \\ &\quad \alpha_{n-1}(\hat{F}_E^p(n-1,j) - \tilde{F}_E^p(n-1,j))]^2\} \end{aligned} \tag{9}$$

From Eqn. (9), we can see that the distortion $D(n)$ comes from two parts, i.e., the distortion produced by truncating bits in the current enhancement layer frame, denoted as $D_I(n)$, and the distortion propagated from the previous frame due to the leaky prediction, denoted as $D_P(n-1)$. $D_I(n)$ and $D_P(n-1)$ are defined as

$$D_I(n) = E\{[F_E(n,i) - \alpha_{n-1}\hat{F}_E^p(n-1,j) - \tilde{e}_E(n,i)]^2\}, \tag{10}$$

$$D_P(n-1) = E\{[\hat{F}_E^p(n-1,i) - \tilde{F}_E^p(n-1,i)]^2\}. \tag{11}$$

Similar to [10], we assume that

$$E\{[\hat{F}_E^p(n-1,j) - \tilde{F}_E^p(n-1,j)]^2\} = \rho_{n-1}D_P(n-1), \tag{12}$$

where ρ_{n-1} is a constant describing the motion randomness of the video scene. Through extensive experiments, we found that $D_I(n)$ and $D_P(n-1)$ are not independent, and $D(n)$ can be approximately composed as

$$D(n) = D_I(n) + \alpha_{n-1}^2\rho_{n-1}D_P(n-1) + (a_n + b_n\sqrt{D_I(n)})\sqrt{D_P(n-1)}, \tag{13}$$

where a_n and b_n are the constants, and the third term corresponds to the correlation between $D_I(n)$ and $D_P(n-1)$.

$D_I(n)$ can be calculated using the similar method employed in [7,11], i.e., the linear interpolation technique, which can be performed at the offline stage. It is not easy to compute $D_P(n-1)$ since it depends on the bits allocated to the $(n-1)$-th enhancement layer frame and the bits used for the partial leaky prediction. There are two situations.

(1) The amount of the allocated bits is smaller than that for the partial leaky prediction, i.e., $\tilde{F}_E(n-1,i) = \tilde{F}_E^p(n-1,i)$. In this case, $D(n-1)$ can be approximated as

$$\begin{aligned} D(n-1) &= E\{[F_E(n-1,i) - \tilde{F}_E^p(n-1,i)]^2\} \\ &= E\{[F_E(n-1,i) - \hat{F}_E^p(n-1,i) + \hat{F}_E^p(n-1,i) - \tilde{F}_E^p(n-1,i)]^2\} \\ &= E\{[F_E(n-1,i) - \hat{F}_E^p(n-1,i)]^2\} + E\{[\hat{F}_E^p(n-1,i) - \tilde{F}_E^p(n-1,i)]^2\} \\ &\quad +(c_{n-1} + d_{n-1}\sqrt{E\{[F_E(n-1,i) - \tilde{F}_E^p(n-1,i)]^2\}}) \\ &= E\{[F_E(n-1,i) - \hat{F}_E^p(n-1,i)]^2\} + D_P(n-1) + \\ &\quad (c_{n-1} + d_{n-1}\sqrt{D(n-1)}), \end{aligned} \tag{14}$$

where c_{n-1} and d_{n-1} are two constants. Therefore, $D_P(n-1)$ is calculated as

$$\begin{aligned} D_P(n-1) = D(n-1) - E\{[F_E(n-1,i) - \hat{F}_E^p(n-1,i)]^2\} - \\ (c_{n-1} + d_{n-1}\sqrt{D(n-1)}). \end{aligned} \tag{15}$$

Note that the value of $E\{[F_E(n-1,i) - \hat{F}_E^p(n-1,i)]^2\}$ can be computed in the offline stage.
(2) The amount of the allocated bits is larger than or equal to that for the partial leaky prediction. Thus, $D_P(n-1)$ can be calculated as

$$\begin{aligned} D_P(n-1) &= E\{[\alpha_{n-2}\hat{F}_E^p(n-2,j) - \alpha_{n-2}\tilde{F}_E^p(n-2,j)]^2\} \\ &= \alpha_{n-2}^2\rho_{n-2}D_P(n-2). \end{aligned} \tag{16}$$

4 Experimental Results

In this section, we conduct experiments to test the accuracy of the proposed R-D model. The experiments are performed on two QCIF format video sequences. The first video sequence is the Foreman sequence with 300 frames, which contains large facial movements and camera panning at the end. The second one is the Akiyo sequence with 300 frames, which contains low activity (slow motion). Both video sequences are coded at 10 fps. The base layers are coded at 32 kbps for the Foreman sequence and 16 kbps for the Akiyo sequence, respectively. In the enhancement layer encoding of the Foreman sequence, for each frame, an amount of 16 kbits is used in the partial prediction with a leaky factor of 0.7. For the Akiyo sequence, an amount of 11 kbits is used in the partial prediction with a leaky factor of 0.8 for each frame.

Figures 1 and 2 show the results of the distortion estimation in the case of having constant channel bandwidth. The difference between the estimated distortion values and the actual distortion values is very small, maximally 1.55 and 0.9 in Figs. 1 and 2, respectively. Fig. 3 shows the results at a time-variable channel condition. Again, it can be observed that the estimation error is almost not noticeable. Note that the estimated distortions are calculated completely

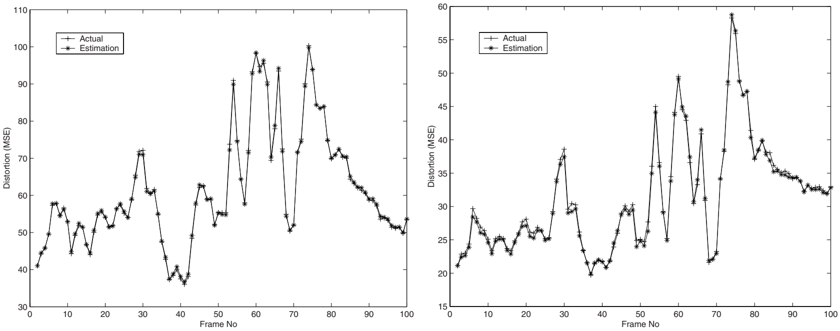


Fig. 1. Distortion estimation results for QCIF Foreman video sequence coded at a fixed total bandwidth. Left: QCIF Foreman at 64 kbps. Right: QCIF Foreman at 128 kbps.

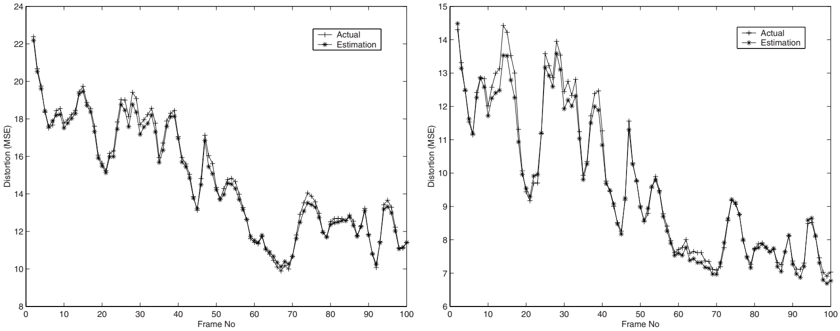


Fig. 2. Distortion estimation results for QCIF Akiyo video sequence coded at a fixed total bandwidth. Left: QCIF Akiyo at 48 kbps. Right: QCIF Akiyo at 80 kbps.

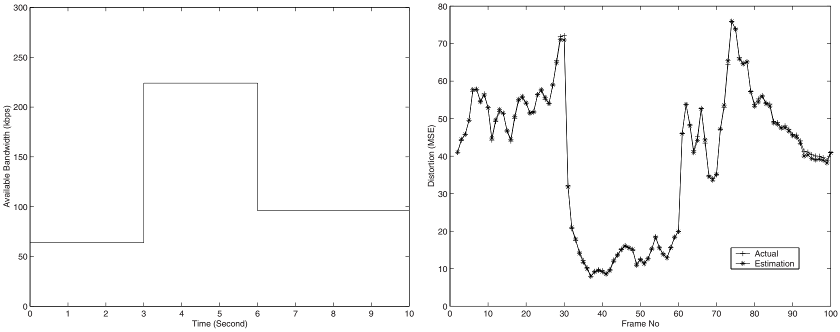


Fig. 3. Distortion estimation results for QCIF Foreman video sequence coded at a variable total bandwidth. Left: Time-variable available bandwidth. Right: Distortion estimation results.

based on the proposed R-D model without any update from the actual distortion values. These experimental results demonstrate the accuracy of our proposed R-D model.

5 Conclusions

In this paper, we have proposed an accurate R-D model for the leaky prediction based FGS video coding. Our proposed R-D model have taken into consideration both the distortion introduced in the current frame and the propagated distortion due to leaky prediction. The experimental results have demonstrated the accuracy of the proposed R-D model. Our future work will focus on applying the proposed R-D model for the optimal bit allocation in L-FGS video coding.

References

1. Li, W.: Overview of fine granularity scalability in MPEG-4 video standard. IEEE Trans. on Circuits and Systems for Video Technology **11** (2001) 301–317
2. Wu, F., Li, S., Zhang, Y.Q.: A framework of efficient progressive fine granularity scalable video coding. IEEE Trans. on Circuits and Systems for Video Technology **11** (2001) 332–344
3. van der Schaar, M., Radha, H.: Motion-compensation based find-granular scalability (MC-FGS). In: ISO/IEC JTC1/SC29/WG11, MPEG00/M6475. (2000)
4. Huang, H., Wang, C., Chiang, T.: A robust fine granularity scalability using trellis-based predictive leak. IEEE Trans. on Circuits and Systems for Video Technology **12** (2002) 372–385
5. Liu, Y., Li, Z., Salama, P., Delp, E.: A discussion of leaky prediction based scalable coding. In: IEEE ICME 2003. (2003) 565 – 568
6. Sangeun Han; Girod, B.: Robust and efficient scalable video coding with leaky prediction. In: IEEE ICIP 2002. (2002) II–41 – II–44
7. Zhang, X.M., Vetro, A., Shi, Y.Q., Sun, H.: Constant quality constrained rate allocation for FGS-coded video. IEEE Trans. on Circuits and Systems for Video Technology **13** (2003) 121–130
8. Dai, M., Loguinov, D., Radha, H.: Analysis of rate-distortion functions and congestion control in scalable internet video streaming. In: ACM NOSSDAV '03. (2003) 60 – 69
9. Wang, Q., Xiong, Z., Wu, F., Li, S.: Optimal rate allocation for progressive fine granularity scalable video coding. IEEE Signal Processing Letters **9** (2002) 33 – 39
10. He, Z., Cai, J., Chen, C.W.: Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding. IEEE Trans. on Circuits and Systems for Video Technology **12** (2002) 511–523
11. Wu, J., Cai, J.: Wireless FGS video transmission using adaptive mode selection and unequal error protection. In: SPIE VCIP 2004. (2004) 1305–1316

JPEG Quantization Table Design for Photos with Face in Wireless Handset

Gu-Min Jeong[1], Jun-Ho Kang[2], Yong-Su Mun[2], and Doo-Hee Jung[3]

[1] SK Telecom, 99, Seorin-dong, Jongro-gu, Seoul, Korea
znco1004@freechal.com
[2] NeoMtel Cooperation, 7th floor 726-11,
Yeoksam-dong, Kangnam-ku, Seoul, Korea
{jhkang, ysmun}@neomtel.com
[3] Dept. of Electrical and Electronics Engineering, Korea Polytechnic University,
2121, Jungwang-Dong, Shihung-City, Kyunggi-Do, 429-793, Korea
doohee@kpu.ac.kr

Abstract. In this paper, a design technique of JPEG quantization table is proposed for photos with face whose size is 128× 128 in wireless handset. The small size images may have different characteristics from the large images used in PC. Also, the photos with face have their own characteristics. From these two facts, new quantization table design is proposed in this paper. The quantization tables are derived from R-D optimization for the photos with face and it is shown that the proposed quantization tables have good performances for size and quality. In R-D optimization, the obtained quantization table is image-specific and cannot be applied to other images. In the proposed method, the quantization tables are obtained from various photo samples and final table is gotten as average of them. This makes the final quantization tables applied to other photos with face. Also, it is possible to control the quality factor based on the interpolation of high quality quantization table and middle quality quantization table. Simulation results show that the obtained quantization table makes the compressed file size small and the image quality improved in general cases.

1 Introduction

Multimedia technology is essential for the saving and transmission of information and is widely used according to the development of communications. For image compression technology, there are various standards. Especially in still image coding, JPEG [1] becomes a *de facto* standard. JPEG is adopted various areas such as PC, digital camera, broadcasting and so on. According to the development of wireless communication, JPEG is utilized in wallpaper or photo services. JPEG is also a standard codec for 3GPP [5] and 3GPP2 [6]. Almost of handsets will use JPEG and the usage of JPEG will increase in proportion to the development of camera module in handset. Due to some restrictions of handset such as memory limitation, low CPU processing power, narrow network bandwidth, implementation of JPEG player in handset is different from that of

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 681–688, 2004.

PC. JPEG player in handset must be fast and the contents size should be small. Low bit rate coding should be made for wireless internet service. More than the porting environment such as memory and CPU, the characteristics of image with small size in handset is different from that of PC. The LCD size is small in handset and the contents provider should resize the contents according to the LCD. This makes the DCT frequency distribution in 8×8 block change. Roughly speaking, the contents now serviced in handset have more high frequency components than that of PC. Considering these characteristics of images in handset, a new quantization table design method is proposed in this paper especially for the photos with face. An optimal quantization table can be obtained using the R-D optimization technique [2] for one image. In R-D optimization, one quantization table is obtained per one image and it is difficult to apply it to other images. In this paper, the images are restricted to the photos with face for a size 128×128. The quantization tables are calculated for various photo samples with R-D optimization. The final tables for quality 75 and quality 50 are obtained as averages of those quantization tables respectively. Considering these two tables corresponding to quality 75 and quality 50, the quality control can be done with interpolation. The final quantization tables are applied to the JPEG player for photos with face. The test results with the proposed method applied in IJG JPEG source show that the compressed file size is smaller and the image quality is better than the results with the default JPEG quantization table in general cases. The remainder of this paper is organized as follows. In Section 2, the characteristics of handset images and R-D optimization are briefly summarized. In Section 3, the quantization table selection algorithms are proposed. In Section 4, test results are presented and the conclusion follows in Section 5.

2 The Characteristics of Handset Images and R-D Optimization

2.1 The Characteristics of Images in Handset

Usual handsets in Korea have a LCD with a size 128×160. High-end handsets for MPEG 4 services have a 176×200 or 176×220 LCD. In this paper, the LCD size 128×160 is considered. In 128×160 LCD, the wallpaper images have a size with 128×128. In this paper, quantization tables are designed for 128×128. For wallpaper of handset, the image codecs such as SIS [7], JPEG, MPEG etc. are used in Korea. For photo images, JPEG is mostly used since the user takes a picture with JPEG in handset and sends the image with JPEG. The characteristics for 128×128 images in handset may be different from those of PC images. Based on these differences of characteristics, some quantization table may have a better performance than the default quantization table in JPEG. If it is assumed that the images in PC have a size 512×512, the images in handset have a size with 128×128 and the horizontal size and vertical size are 1/4 of PC images respectively. Let us consider the photo image Fig. 1. In Fig. 1, the original images are displayed for 512×512 and 128×128 respectively and the left-lower 8×8 blocks

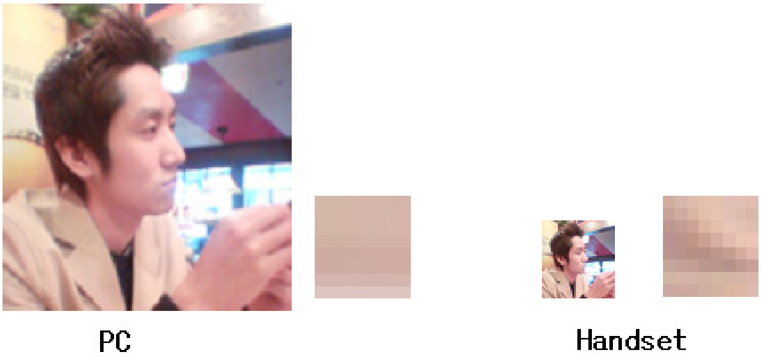


Fig. 1. Display in PC / handset and the left lower 8×8 blocks

are displayed. As shown in left-lower blocks, if we resize the 512×512 images into 128×128 images, the frequency distributions in the images can change for 8×8 block. That is, the high frequency components may increase in 8×8 block. In JPEG compression, quantization table makes high frequency coefficients nearly zero. However as in Fig. 1, 128×128 images may have more high frequency components than 512×512 images. This means that the default JPEG quantization table may perform bad for small images and some quantization tables which compensate for high frequency may have better performance. Also, the photos with face consist of background and face as Fig. 1. The frequency distribution of them is different from that of general images.

In this paper, considering these points, quantization tables are designed for photos with face in handset with 128×128 LCD. The implementation is based on IJG JPEG open source [3]. The quantization table design is done for Arai methods [4] in IJG source.

2.2 R-D Optimization

R-D optimization is an algorithm to efficiently optimize rate-quality. In R-D optimization, objective function is given as

$$D(Q) + \lambda R(Q) \tag{1}$$

where λ is an Lagrange multiplier, D, R, Q denote the distortion, rate, quantization table, respectively. The objective function (1) is minimized according to λ and quantization table Q. The R-D optimization can be summarized as follows:
(**1**) Set $\lambda = 0$.
(**2**) Increase the each component of quantization table and find the optimal Q which minimizes the R-D function (1).
(**3**) Increase λ by δ and repeat Step 2.

(**4**) Repeat Step 3 until λ reaches λ_{max}.
(**5**) Plot R-D curve according to the variation of λ and find λ which is tangential to that curve with λ slope. At this time, the λ is the optimal λ and the quantization table is an optimal quantization table.

3 Quantization Table Design

3.1 Quantization Table Selection Using R-D Optimization

R-D optimization is an algorithm to efficiently optimize rate-quality tradeoffs in an image-specific way. However, to obtain a quantization table, R-D optimization is image-specific and it requires too much time. Due to the timing issues, it cannot be applicable to real services. Even for the cases encoding in PC, it takes too much time for obtaining quantization table and it cannot be utilized. As shown in Section 2, the images serviced in handset have characteristics different from those of PC. The default JPEG quantization table may perform bad for small images and some quantization tables which compensate for high frequency may have better performance. Since R-D optimization is an image-specific way, for the fast processing, we consider only photos with face for image size 128×128. It is assumed that the photos with face have similar characteristics each other.

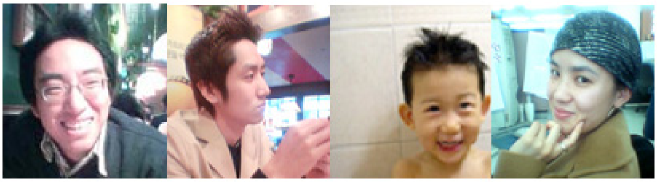


Fig. 2. Examples of typical images

In this paper, for low bit rate coding in wireless handset, the quantization table design is based on R-D optimization and the quality control is also done with two quantization tables Q75 and Q50. For low bit rate coding, the proposed algorithms in this paper are as follows:
(**1**) The photos with face now serviced in wireless handset are searched. Typical 100 images are selected. Some typical images are shown in Fig. 2.
(**2**) Each image is compressed with image quality 75 (50).
(**3**) For each image, find $R_{min}75(R_{min}50)$, $R_{max}75(R_{max}50)$, respectively.
(**4**) 4. For each image, calculate $(R_{min}75 + R_{min}50)/2$ and find quantization tables Q75 (Q50) corresponding to $(R_{min}75 + R_{min}50)/2$ as shown in Fig. 3.
(**5**) 5. The final quantization table Q75 (Q50) for quality 75 (50) as an average of the quantization tables in step 4.

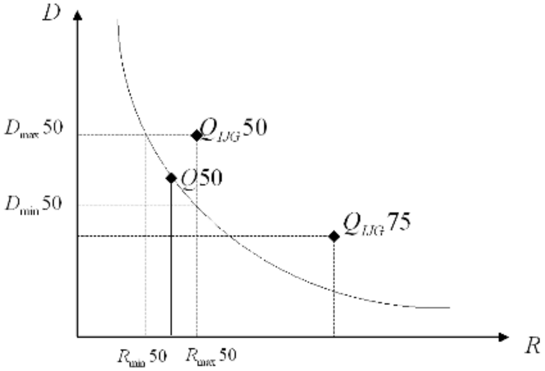


Fig. 3. Quantization table selection for image quality 75 and 50 for each image

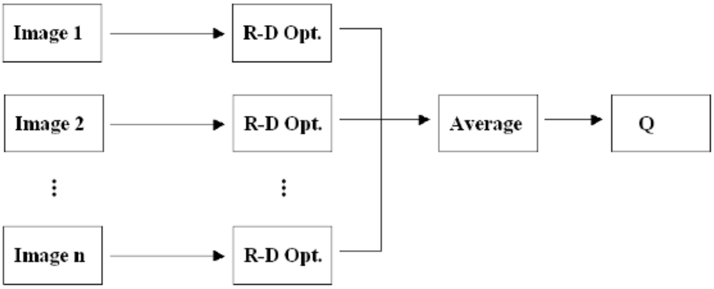


Fig. 4. Final quantization table selection for image quality 75 and 50

Fig. 3 shows the quantization table selection algorithms briefly. After quantization table is selected for one image, the final quantization table is calculated as an average as in Fig. 4.

The images in Fig. 2 are the examples of the typical images. The quantization tables are calculated from the images in Fig. 2 using the proposed algorithms. Table 1 shows the final quantization table Q 75 for quality 75 and Table 2 shows quantization table Q 50 for quality 50.

In JPEG compression, the compression gain is obtained from the loss of high frequency and in a default JPEG quantization table the values becomes larger for high frequencies. As shown in Table 1 and Table 2, the values of the proposed quantization tables are uniform for all frequencies. It may be because the images serviced in handset may have more high frequency components than the images in PC.

Table 1. IJG's and proposed quantization tables for quality 75

8	6	5	8	12	20	26	31
6	6	7	10	13	29	30	28
7	7	8	12	20	29	35	28
7	9	11	15	26	44	40	31
9	11	19	28	34	55	52	39
12	18	28	32	41	52	57	46
25	32	39	44	52	61	60	51
36	46	48	49	56	50	52	50

11	10	9	9	7	11	9	9
10	9	8	8	9	11	9	11
10	10	10	9	10	11	10	11
10	10	8	12	11	12	9	11
9	8	9	11	11	12	9	9
8	9	10	10	10	11	9	11
9	9	9	10	10	10	10	11
13	11	12	11	8	11	15	13

Table 2. IJG's and proposed quantization tables for quality 50

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

37	28	29	26	23	26	23	25
33	34	25	32	31	34	22	26
36	25	36	27	41	27	29	25
30	43	25	40	32	41	23	25
43	18	42	15	37	22	24	22
22	38	20	39	30	31	17	20
40	18	35	27	40	23	26	23
31	33	31	46	22	30	21	31

3.2 Quality Control with Interpolation

In JPEG compression, the quality control is done using quality factor. It is possible to control the image quality and the file size using quality factor. In wireless handset, it is also important to control the quality factor due to the low bit rate coding. In this paper, the quality control is done from quality 50 to quality 75 based on the interpolation of quantization table between Q75 and Q50.

4 Results

The calculated Q 75 and Q 50 are implemented in IJG JPEG source. The test is divided into two parts. The first test is for the 100 typical images. Though Q75 and Q50 are obtained from those images, Q75 and Q 50 may not match these images since the quantization tables are averages. In Test 1, the quality and size are checked with Q75, Q50 and Q60. It should be noted that Q 60 is an interpolation from Q75 and Q50. The second test is for general images which are not included in the typical images.

4.1 Tests with Typical Images

Table 3 shows the results for the typical 100 images. At quality 75, in 76 images, quality is better and size is smaller than the results of IJG. In some images, size or quality is bad. However, there is no image in which both quality and size is bad.

Table 3. Results for 100 typical images

Rate	All good	Size bad	PSNR bad	All bad
Quality 75	76	10	14	0
Quality 60	60	22	18	0
Quality 50	46	22	32	0

4.2 Tests with Other Images

Test 2 is done for other images not including in typical images. Fig. 5 shows the images used in this test.

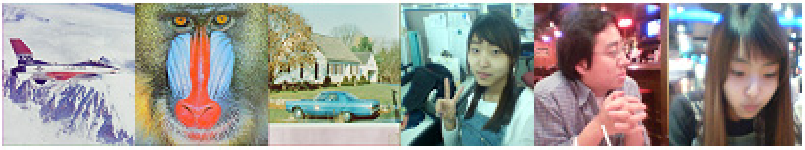


Fig. 5. Images for test 2

As shown in Fig. 5, The former 3 images are general test images such as jet plane, barboon, car. The other 3 images are photos with face.

As in Table 4, the results for general test images (image 1, image 2 and image 3) are not so good. However, the results for photos with face (image 4, image 5 and image 6) are similar to that of test 1. For all images, the compressed image size is smaller that that of IJG. The image quality is better in image 4 and 5. In image 6, the image quality is a little bad.

In the results of test 1 and test 2, the proposed quantization table Q 75 and Q 50 performs better in size and quality than IJG quantization table for the photos with face. Generally R-D optimization technique consumes too much time, while the proposed method calculates the quantization table beforehand and the processing time is fast.

Table 4. Results for other images

	Q75				Q50			
	Size (bytes)		PSNR (dB)		Size (bytes)		PSNR (dB)	
Image 1	4085	4023	28.3721	29.2781	2707	3050	26.4979	27.8523
Image 2	5570	6034	25.2746	26.3322	3563	4623	23.4113	24.9886
Image 3	4643	4750	27.0398	27.8950	3105	3625	25.2871	26.5651
Image 4	4081	3725	30.8251	31.0861	2768	2644	29.0674	29.2420
Image 5	4021	3619	30.1237	30.2451	2714	2558	28.5917	28.5945
Image 6	3211	2566	30.8407	30.5631	2192	1773	29.4386	28.9465

5 Conclusion

In this paper, we have proposed a quantization table design technique for photos with face in wireless handset. For the images with size 128×128, new quantization table is proposed for the improvement of image quality and compression ratio. The photos with face whose size is 128×128 have their own characteristics different from general images in PC. Considering these characteristics, new quantization table is designed. The proposed method is derived from the well-known R-D optimization. Using R-D optimization, though the image specific quantization tables can be obtained, the all processing takes too much time and it cannot be applicable to real implementations. In the proposed method, new quantization tables are acquired beforehand from the typical images and applied to the JPEG player. Thus, the proposed method can achieve good performance in the sense of fast processing, small size and good image quality. To get the quantization tables, at first step, the image specific quantization tables corresponding to quality 75 and quality 50 are calculated. Next, the final quantization table Q75 and Q50 are obtained as an average of those tables respectively. Test results show that the proposed quantization tables have better performance than those of JPEG quantization tables in the sense of image size and image quality for photos with face whose size is 128×128.

References

1. Wallace, G. K.: The JPEG Still-Picture Compression Standard. Commun. ACM, Vol. 34, Apr. 1991 30–44
2. Crouse, M., Ramchandran, K.: Joint Thresholding and Quantizer Selection for Transform Image Coding : Entropy-Constrained Analysis and Application to Baseline JPEG. IEEE Trans. on Image Processing, Vol. 6, No. 2, Feb, 1997 285–297
3. Independent JPEG Group, http://www.ijg.org
4. Arai, Y., Agui, T., Nakajima, N.: A New DCT-SQ Scheme for Images. IEICE, Vol. E71, 1988 1095–1097
5. 3GPP, http://www.3gpp.org
6. 3GPP2, http://www.3gpp2.org
7. Neomtel, http://www.neomtel.com

Effective Drift Reduction Technique for Reduced Bit-Rate Video Adaptation

June-Sok Lee[1], Goo-Rak Kwon[1], Jae-Won Kim[1],
Jae-Yong Lee[2], and Sung-Jea Ko[1]

[1] Department of Electronics Engineering, Korea University, Seoul, Korea
[2] Nextreaming Corporation, Seoul, Korea
Tel: +82-2-3290-3228
sjko@dali.korea.ac.kr

Abstract. Video transcoding to the low bitrate creates drift error caused by quantization, motion vector mapping, and texture downsizing. In this paper, we propose a drift reduction technique for spatial resolution reduction transcoding. The proposed technique uses *adaptive motion mapping and refinement* for avoiding the generation of the drift error and *forced skipping of macroblock encoding* for terminating the propagation of drift error. Experimental results show that proposed technique can preserve image quality while transcoding to the low bitrate.

Keywords: Video adaptation, Transcoding, Drift error reduction.

1 Introduction

Recently, MPEG initiated a standardization for the digital item adaptation on the various terminal capabilities, network conditions, and preferences of user [1]. In this aspect, the video transcoding is an efficient mechanism to deliver visual contents to a variety of users who have different network conditions or terminal devices with different display capabilities.

Video transcoding is a process to convert one format to another with different properties which include frame size, frame rate, bit rate, and syntax [2]. Especially, the spatial resolution reduction (SRR) transcoding is a useful video transcoding method when the target bitrate is low or the clients have small display capabilities. The major problem of the SRR video transcoding is the picture quality degradation that comes from drift error. The drift error refers to the continuous decrease in picture quality that occurs when the reconstructed pictures in the encoder and the decoder are not exactly the same. An analysis provided in [3] identifies the sources of drift error as the mismatch of quantizer and motion vector (MV) between source decoder and target encoder. Especially, MV mapping (MVM) used to obtain the downscaled MV's is the prominent source of drift error in the SRR transcoding since the MV's obtained in the inaccurate MVM process are used for the motion compensation at the target encoder.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 689–696, 2004.

Many researches have been conducted on video transcoding to remove the drift error [3]-[6]. The drift error can be eliminated by using the intra refresh method which interleaves intra frames or intra macroblock's (MB's) in the bitstream at certain intervals [3]. However, the intra refresh method is not proper to the low bitrate (LBR) video transcoding since the insertion of intra coded MB's or frames increases the bitrate of the encoded bitstream. In [4], an adaptive MV resampling method is proposed to compensate the drift error. The MV resampling method is a useful technique to enhance the picture quality without the use of the intra coded information. However, the performance of the MV resampling method highly depends on the accuracy of the resampled MV since the inaccurately resampled MV leads to the degradation of transcoded picture quality. In [5]-[6], the MV refinement (MVR) method is used to solve the drift error problem by improving the accuracy of the MV.

In this paper, we propose a drift reduction (DR) technique for the SRR transcoding. The proposed DR technique uses *adaptive motion mapping and refinement* (AMMR) and *forced skipping of MB encoding* (FSMBE). The AMMR method can significantly reduce the drift error by improving the accuracy of the downscaled MV's. The FSMBE method is used to terminate the propagation of the drift error by using the MB's encoded with the skip mode (or skipped MB's). Unlike the intra refresh method, the FSMBE method does not increase the bitrate of the transcoded bitstream.

The rest of the paper is organized as follows. In Section 2, the proposed DR technique is described in detail. The experimental results and conclusions are given in Section 3.

2 Proposed Drift Reduction Technique

Fig. 1 shows the architecture of the proposed video transcoding system. The proposed transcoder reduces the spatial resolution of the input sequence by a factor of 2 and generates the MV for the downscaled frame. The transcoded information is transferred to the target encoder and encoded to generate the compressed bitstream. In the proposed transcoder, the AMMR unit selects the precise MV for the downscaled MB and refines the selected MV by using an adaptively changing search range in the downscaled frame. The FSMBE unit selects the MB's to be encoded with a skip mode to terminate the propagation of drift error. The skipped MB's are served as the intra coded MB's at the decoder since the decoder copies the collocated MB's in the reference frame to reconstruct the skipped MB's. Next, we describe the proposed DR technique in detail.

2.1 Adaptive Motion Mapping and Refinement (AMMR)

In SRR transcoding, a group of four 16×16 MB's in the original frame corresponds to one 16×16 MB in downsized frame. Thus, MVM is required to allocate an MV for the downscaled MB. However, the simple MVM method

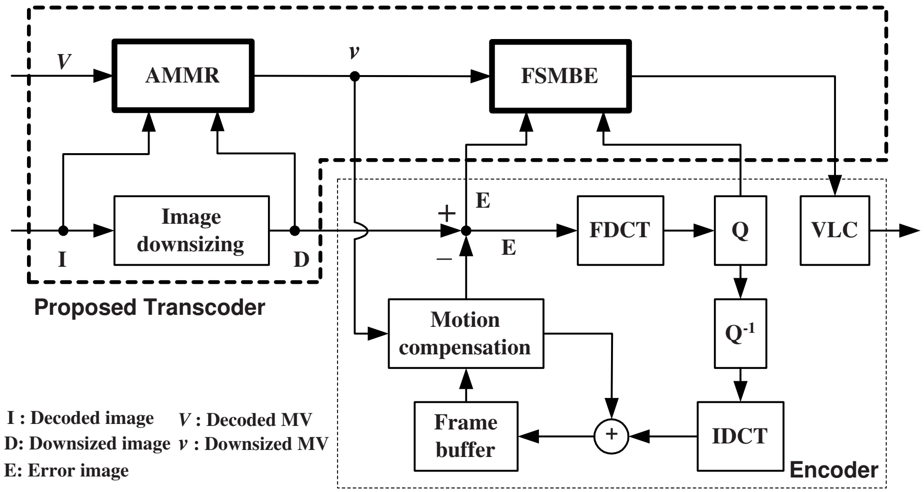


Fig. 1. Proposed video transcoding architecture

such as MVM by averaging or median in [5] may generate inaccurate MV's due to the variety of MV and MB types. For example, in Fig. 2 (a), the top left MB has two MV's since the MB is encoded with the field MV type, and the top right MB has 8×8 MV type. The bottom left MB has two MV's; the forward MV and the backward MV. The bottom right MB is the intra coded MB with no MV. Therefore, various MB and MV types must be considered in the MVM process to allocate an MV for the downscaled MB.

For the the downscaled MB, the proposed AMMR method selects an MV that produces the minimum prediction error. The AMMR method consists of *representative MV (RMV) extraction*, *base MV (BMV) extraction*, and *BMV refinement* processes as shown in Fig. 2. In the first step, RMV extraction is performed to make a single MV for each MB in the original frame since the MB

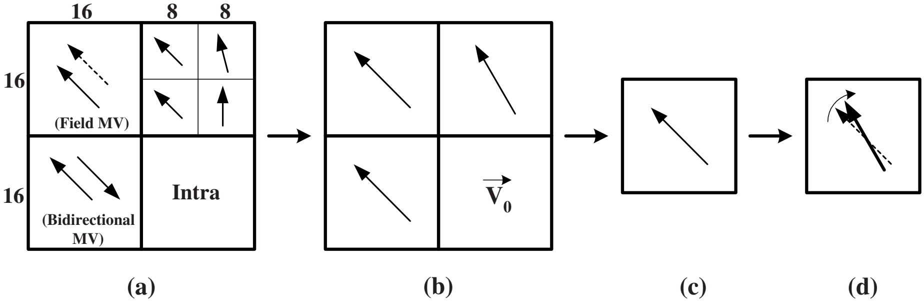


Fig. 2. Adaptive motion mapping and refinement method (a) example of MB & MV types, (b) RMV extraction, (c) BMV extraction, (d) BMV refinement

may have multiple MV's or no MV as shown in Fig. 2 (a). The RMV is obtained by selecting an MV minimizing the prediction error among the multiple MV's of the MB, i.e.,

$$\overrightarrow{\mathrm{V}}^{k}_{RMV} = \arg\min_{\overrightarrow{\mathrm{V}} \in S} \left(\frac{1}{N} \sum_{j=0}^{N-1} \left| P_j^k - R_j^k(\overrightarrow{\mathrm{V}}) \right|^2 \right), \tag{1}$$

where $\overrightarrow{V}^{k}_{RMV}$ is the RMV of the k^{th} MB, S is the set of MV candidates, N is the number of pixels in the MB, the P_j^k is the pixel in the k^{th} MB, and $R_j^k(\overrightarrow{V})$ is the pixel in the reference MB predicted with $\overrightarrow{V}$. Table 1 summarizes the MV candidates for different MB and MV types. The zero MV ($\overrightarrow{V}_0$) is used for the RMV of the intra-coded and skipped MB's since these MB's do not take MV's. For the forward MB with the field MV type, the RMV is selected from the set $S = \{\overrightarrow{V}_0, \overrightarrow{V}_t, \overrightarrow{V}_b, \overrightarrow{V}_{avg}\}$, where $\overrightarrow{V}_t$ is the top-field MV, $\overrightarrow{V}_b$ is the bottom-field MV, and $\overrightarrow{V}_{avg}$ is the average of all the candidate MV's in the set. For the bidirectional MB with frame MV type, the RMV is selected from the set $S = \{\overrightarrow{V}_0, \overrightarrow{V}_{fwd}, \overrightarrow{V}_{bwd}, \overrightarrow{V}_{avg}\}$, where $\overrightarrow{V}_{fwd}$ and $\overrightarrow{V}_{bwd}$ represent the forward and backward MV's, respectively. For the backward MB type, the MV candidates are first reversed and then the RMV is selected from the reversed candidate set.

Table 1. Sets of MV candidates for different MB types

MB type	Frame MV type	Field MV type
Skipped/Intra	$\overrightarrow{V}_0$	$\overrightarrow{V}_0$
Forward	$\overrightarrow{V}_0, \overrightarrow{V}_{fwd}$	$\overrightarrow{V}_0, \overrightarrow{V}_t, \overrightarrow{V}_b, \overrightarrow{V}_{avg}$
Backward	$\overrightarrow{V}_0, -\overrightarrow{V}_{bwd}$	$\overrightarrow{V}_0, -\overrightarrow{V}_t, -\overrightarrow{V}_b, -\overrightarrow{V}_{avg}$
Bidirectional	$\overrightarrow{V}_0, \overrightarrow{V}_{fwd}, \overrightarrow{V}_{bwd}, \overrightarrow{V}_{avg}$	$\overrightarrow{V}_0, \overrightarrow{V}_t, \overrightarrow{V}_b, \overrightarrow{V}_{avg}$

In the second step, the BMV is extracted for the downscaled MB as shown in Fig. 2 (c). The BMV is the MV that produces the minimum prediction error among the four RMV's, and given by

$$\overrightarrow{\mathrm{V}}^{k}_{BMV} = \arg\min_{\overrightarrow{\mathrm{V}} \in S_R} \left(\frac{1}{N} \sum_{j=0}^{N-1} \left| \hat{P}_j^k - \hat{R}_j^k(\overrightarrow{\mathrm{V}}) \right|^2 \right), \tag{2}$$

where $\overrightarrow{V}^{k}_{BMV}$ is the BMV of the k^{th} MB in the downsized frame, S_R is the set of the RMV's, the $\hat{P}_j^k$ is the pixel of the k^{th} MB in the downsized frame, and $\hat{R}_j^k(\overrightarrow{V})$ is the pixel in the reference MB predicted with $\overrightarrow{V}$ in the downsized frame.

Finally, the BMV is refined to achieve the enhanced coding efficiency and picture quality of the transcoded frame. To reduce the computation, we refine the BMV only for the MB's with large mean square error (MSE) as follows:

$$\begin{cases} \text{BMV refinement}, & \varepsilon_i > T_u \\ \text{No refinement}, & otherwise \end{cases}, \tag{3}$$

where ε_i is the MSE of the i^{th} MB and T_u is a threshold for the BMV refinement. We set T_u to be equal to the average of all ε_i's. For the selected BMV to be refined, the search range is calculated as follows:

$$S_k = \left\lfloor \frac{\varepsilon_i}{\varepsilon_{\max}} \times S_{max} + 0.5 \right\rfloor, \tag{4}$$

where S_k is the search range to refine the BMV of the k^{th} MB, $\varepsilon_{\max}$ is the maximum MSE, and S_{max} is the maximum search range to be determined empirically. Figure 3 shows an example of the MSE of each MB for a certain frame of the "Foreman" sequence. The proposed AMMR method refines the BMV only for MB's with MSE greater than the T_u.

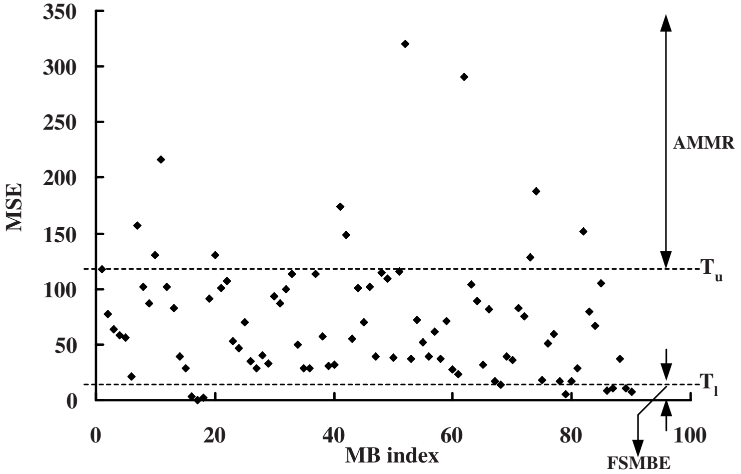


Fig. 3. Thresholds for AMMR and FSMBE

2.2 Forced Skipping of MB Encoding (FSMBE)

As described in Section 1, the intra refreshing is an useful method to terminate the propagation of drift errors. Under the low bitrate constraints, however, the intra refreshing method degrades the quality of transcoded video since too many bits are consumed by the intra coded MB's [3]. To solve this problem, we propose the FSMBE method that can remove the drift error propagation without increasing the bitrate of the transcoded bitstream.

In the transcoding process, the FSMBE method assigns the skip mode to MB's with small MSE. At the decoder, the skipped MB is copied from the collated MB in the reference frame. Since the copied MB is served as the intra coded MB, the drift error can be removed. As shown in Fig. 3, the FSMBE method skips the MB encoding whenever the MSE is smaller than threshold T_l given by

$$T_l = \frac{1}{M} \sum_{k=1}^{M} \{\varepsilon_k \mid \varepsilon_k < \varepsilon_f \text{ and } \vec{\upsilon}_k = 0\}, \tag{5}$$

where M is the number of MB's whose $\varepsilon_k < \varepsilon_f$ and $\vec{\upsilon}_k = 0$. Note that the skipped MB has $\vec{\upsilon}_k = 0$ since the MB is copied from the collocated MB in the reference frame at the decoder.

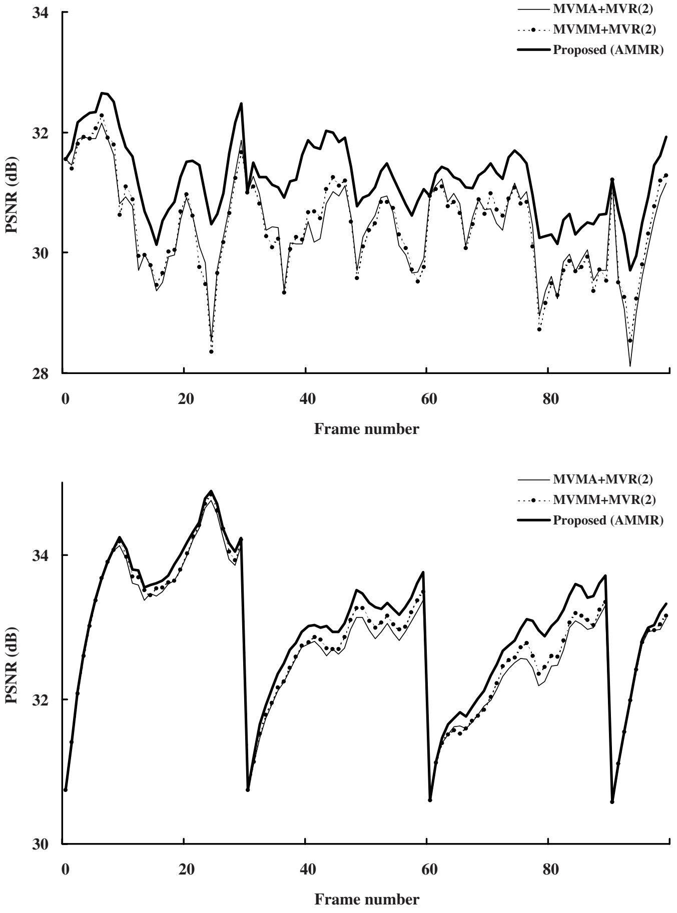
As in the video coding standards such as MPEG-4 and H.263, the FSMBE allocates only one bit for the skipped MB. That is, no bit is consumed to encode the MV and DCT coefficients for the skipped MB. The FSMBE method can allocate the extra bits saved by the aforementioned skipped mode for the other non-skipped MB's in the frame, achieving the enhancement of the transcoded picture quality.

3 Experimental Results and Conclusions

The SRR transcoding from MPEG-2 to H.263 is used to evaluate the performance of the proposed DR technique. The four test sequences, "Foreman", "Stefan", "Table tennis", and "Silent", with CIF resolution (352×288) are encoded to produce the MPEG-2 bitstreams of 1.15 M bps with 30 frames in GOP ($N = 30$) and 3 frames in sub group ($M = 3$). The input MPEG-2 bitstream is transcoded to H.263 bitstream of 64 Kbps with 15 fps.

The performance of the AMMR method is compared with that of the MVM by averaging (MVMA) and MVM by median (MVMM) methods in [5]. Fig. 4 shows the transcoded picture quality in terms of PSNR. For the MVR process of MVMA and MVMM, the full-scale motion estimation is applied with the search range of ±2. As shown in Fig. 4, the proposed AMMR method achieves higher picture quality since it provides more accurate MV's for the SRR transcoding than the MVMA and MVMM methods.

The performance of the proposed FSMBE method is compared with that of the intra refresh (IR) method in the H.263 video coding standard [8]. In the IR method, each MB has a counter that is increased if the MB is encoded in the inter coding mode. If the counter reaches a threshold $T = 1/\beta$ where β is the intra refresh rate, the MB is encoded with the intra coding mode and the counter is reset to zero. If $\beta = 1$, all the MB's are encoded with the intra coding mode. Table 2 shows that the proposed FSMBE method achieves higher picture quality with lower bitrate than the IR method with various β's. The IR method



(a) (b)

Fig. 4. Comparison of PSNR corresponding to the test sequence (a) Foreman, (b) Silent

Table 2. The performance of FSMBE (PSNR [dB], Bitrate [Kbps])

	IR ($\beta = 0.1$)		IR ($\beta = 0.01$)		IR ($\beta = 0.002$)		FSMBE	
Test sequence	dB	Kbps	dB	Kbps	dB	Kbps	dB	Kbps
Foreman	29.3	8.8	29.9	8.8	29.9	8.8	30.1	8.7
Stefan	22.7	11.3	22.8	10.9	22.8	10.9	23.2	10.4
Table tennis	30.2	8.9	30.7	8.9	30.7	8.9	31.0	8.8
Silent	32.2	8.6	32.9	8.6	32.9	8.6	33.0	8.5

shows the lowest PSNR and highest bitrate at $\beta = 0.1$ since the IR method consumes many bits for the intra coded MB's under the low bitrate constraint. Even with the very small intra refresh rate, e.g., $\beta = 0.002$, the IR method exhibits lower picture quality and higher bitrate as compared with the proposed FSMBE method.

From the experimental results, it is seen that the proposed DR technique is the useful method to remove the drift error with the improved picture quality in the SRR transcoding. The proposed AMMR method can avoid the generation of the drift error. The FSMBE method terminates the propagation of drift error without increasing of bitrate.

References

1. Vetro, A. (ed.): MPEG-21 requirements for for digital item adaptation. ISO/IEC JTC1/SC29/WG1 N4684, Korea (2002)
2. Vetro, A., Christopoulos, C., Sun, H.: Video transcoding architectures and techniques-An overview. IEEE Signal Processing Magazine (2003) 18–29
3. Yin, P., Vetro, A., Lui, B., Sun, H.: Drift compensation for reduced spatial resolution transcoding. IEEE Trans. Circuits Syst. Video Technol., Vol. 12, (2002) 1009–1020
4. Shen, B., Sethi, I. K., Vasudev, B.: Adaptive motion-vector resampling for compressed video downscaling. IEEE Trans. Circuits Syst. Video Technol., Vol. 9 (1999) 929–936
5. Shanableh, T., Ghanbari, M.: Heterogeneous video transcoding to lower spatio-temperal resolutions and different encoding formats. IEEE Trans. Multimedia, Vol. 2 (2000) 101–110
6. Seo, K., Kim, J.: Motion vector refinement for video downsampling in the DCT domain. IEEE Signal processing letters, Vol. 9 (2002) 356–359
7. Corbera, J. R., Lei, S.: Rate control in DCT video coding for low delay communications. IEEE Trans. Circuits Syst. Video Technol., Vol. 9(1) (1999) 172–185
8. ITU-T Recommendation H.263, Video coding for low bit rate communication, (1998)

On Implementation of a Scalable Wallet-Size Cluster Computing System for Multimedia Applications*

Liang-Teh Lee[1], Kuan-Ching Li[2], Chao-Tung Yang[3], Chia-Ying Tseng[1],
Kang-Yuan Liu[1], and Chih-Hung Hung[4]

[1] Dept. of Computer Science and Engineering, Tatung University, Taipei 104, Taiwan ROC
ltlee@cse.ttu.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] High Performance Computing Laboratory, Dept. of Computer Science and Information Engineering, Tunghai University, Taichung 407, Taiwan ROC
ctyang@mail.thu.edu.tw

[4] Networks and Multimedia Institute, Institute for Information Industry, Taipei 106,
Taiwan ROC, chhung19@iii.org.tw

Abstract. In order to obtain a low-cost, low-power consumption and small sized digital home processing center, especially for multimedia applications, we propose a scalable wallet-size cluster computing system in this paper. Such system uses low-cost, low-power consumption and small sized microprocessors as computing nodes to form a cluster system.

The main reason why we are interested to investigate and build such system is to increase the computing power and to accelerate the H.264 video processing. We analyzed the H.264 decoder scheme and distributed the most heavy-load computations in the decoder scheme to each computing node (or microprocessor) of the cluster system for the processing of signals, in parallel. To run multimedia applications in this cluster system, an efficient real-time video processing is implemented.

Keywords: Multimedia Applications, PC Cluster Systems, Low-Power Consumption

1 Introduction

Multimedia applications typically require heavy load computations for large amount of data. An emerging video coding standard named H.264 or MPEG-4 Part 10 aims to code video sequences at approximately half the bit rate when compared to MPEG-2 at the same quality. Such video coding standard also aims to have significant improvements in coding efficiency, error robustness and network friendliness.

For instance, to copy a video with up to 9GB from a DVD to a CD-ROM, it requires large amount of computing power and time. All the data volume must be reduced to about twelve-th part of its original size, in order to accommodate in the 700MB of

* This research was supported in part by National Science Council, Taiwan, under grant no. NSC92-2213-E-126-006.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 697–704, 2004.

limited storage capacity of a CD-ROM. A data compression of this magnitude for digital video is only possible with the video compression standard MPEG-4. See figure 1 as illustration of this case. Generally, when a DVD title is converted to MPEG-4 format on a single PC, the following steps are performed:

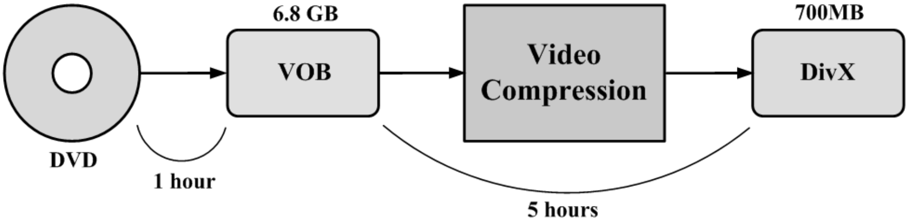


Fig. 1. Typical video conversion - single stream.

In order to increase the computing power and achieve higher performance for video decoding, a full-scale H.264 system decoder using an ARM-based cluster for working collectively is proposed in this paper. Instead of the traditional PC-based cluster system, we use the low cost, low power consumption and small sized ARM chip as computing nodes of a wallet-size cluster system. Additionally, increased computing power of the proposed wallet-size cluster system is achieved with constant advances in IC technology, the problem of performing highly complicated multimedia applications in a low-end processor can perfectly be solved. Additionally, by applying the SoC technology, the cost, power consumption, and the size of the cluster computing system can be reduced significantly. For video coding application, the H.264 standard is selected to be implemented as video decoder for accelerating the decoding speed using our proposed wallet-size cluster system.

This paper is organized as follows. In section 2 it is introduced ARM and H.264 background, while in section 3 is shown the architecture of the proposed wallet-size cluster system. In section 4 is shown how video decoding is processed in this wallet-size cluster system and finally in section 5, some conclusion remarks are discussed.

2 Background

ARM7TDMI is a 32-bit microprocessor [1]. With excellent performance, area and low power consumption properties, it is very attractive for embedded applications. ARM7TDMI is a RISC architecture microprocessor, and its instruction set and related decoding mechanism are much simpler than those show in CISC architectures. This simplicity results in a high instruction throughput and impressive real-time interrupt response from a small die-size, high-performance and cost-effective chip. ARM7TDMI has three-stage pipelining technique, so that all parts of the processing and memory systems can operate continuously. It bus interface is unified, that is, 32-bit data bus carries both instructions and data. The ARM memory interface can allow performance potential and without incurring high costs in the memory system. However, ARM7TDMI does not have an instruction or data cache; and thus, it is mostly used as a controller core rather than for data processing. See 2 for ARM7TDMI diagram.

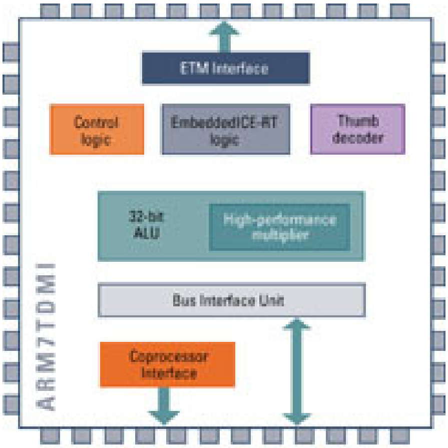


Fig. 2. ARM7TDMI architecture diagram.

During late 2001, ISO/IEC MPEG and ITU-T VCEG decided on a joint venture to wards enhancing standard video coding performance - specifically in the areas where bandwidth and/or storage capacity are limited. This Joint team of both standard organizations was named Join Video Team (JVT). The standard defined by JVT is H.264/MPEG-4 part 10 and at the present time it is referred to as JVT/H.26L/Advanced Video Coding (AVC) [3,10].

H.264 can be widely used in video communication servers in IP network and wireless environment. Its key features when compared to its antecessors are:

- Enhanced motion estimation with variable,
- Enhanced entropy coding,
- Integer block transform,
- Improved in-loop deblocking filter.

The H.264 is an international video coding standard with superior objective and subjective image quality [4,10], reducing average bit rate of 50%, given fixed fidelity when compared to any other standard. The main goals of JVT can be summarized as significant coding efficiency, simple syntax specifications and seamless integration of video coding into all current protocols and multiplex architectures (network friendliness). H.264 meets requirements from the various video applications that aims at supporting video streaming, video conference, over fixed and wireless networks and over different transport protocols, etc.

H.264 has grouped its capabilities into profiles and levels - Baseline, Main and Extended profile. A “profile" is a subset of the entire bit stream of syntax that is specified by the international standard. Within each profile, there are a number of levels designed for a wide range of applications, bit rates, resolutions, qualities and services. A “level" has a specified set of constraints imposed on parameters in a bit stream. It is easier to design a decoder if the profile, level and hence the capabilities are known in advance. Baseline profile can be applied to video conferences and video telephony, the main profile can be applied to broadcast video, and the extended profile can be applied to streaming media.

3 System Architecture

Figure 3 presents the High Speed Parallel Processing System (HSPPS) architecture. We use the embedded RISC microprocessor as computing node to form a cluster computing system [9,6]. A high-speed parallel switching system is also provided to connect each one of computing nodes of the cluster system. Because embedded systems are not efficient and sufficient in storage facilities, a Large Storage Disk (LSD) is setup for each computing node of the cluster system to share and store intermediate/final data. Additionally, we can setup a cluster middleware to support single system image and high availability infrastructure. The proposed wallet-size cluster system provides a parallel programming environment to execute either sequential or parallel applications.

Each computing node of our wallet-size cluster system in current stage of development is an ARM evaluation board, manufactured by MiceTek Company [7]. The uClinux operating system is ported to each board and the Message-Passing Interface (MPI) is installed in each computing node to perform the distributed high-performance processing of the H.264 video decoding.

Figure 4 shows the experimental scheme of the proposed wallet-size cluster system. A personal computer is served as a master node to connect to the Internet for outward communication and manages load balancing among all slave computing nodes. The master node acts as a NIS/NFS server, while the slave computing nodes are NIS/NFS clients with RSH service.

4 Video Decoding in Wallet-Size Cluster System

An abstraction of software environment in the proposed wallet-size cluster system is shown in Figure 4.

The master server node exports its `‘‘user home’’` shared with all slave computing nodes. As described in the previous section, the H.264 standard consists of Profiles and Levels. The Profile specifies a set of algorithmic features and limits, that is supported by all decoders as to that profile. However, users may not require all features provided by H.264. The encoders are not required to make use of any particular set of features supported in a profile. Thus, for any given profile, levels generally correspond to processing power and memory capability on a codec. Each level may support a different picture size - QCIF, CIF, ITU-R 601 (SDTV), HDTV, S-HDTV, D-Cinema and data rate varies from a few tens of kilobits per second (Kbps) to hundreds of megabits per

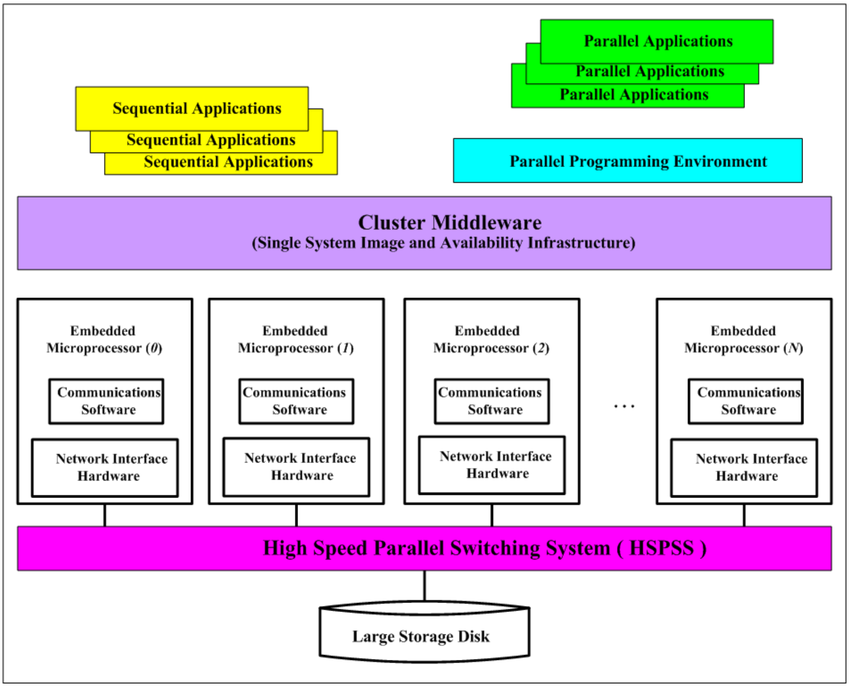

Parallel Applications
Parallel Applications
Parallel Applications
Sequential Applications
Sequential Applications
Sequential Applications
Parallel Programming Environment
Cluster Middleware
(Single System Image and Availability Infrastructure)
Embedded Microprocessor (0)
Communications Software
Network Interface Hardware
Embedded Microprocessor (1)
Communications Software
Network Interface Hardware
Embedded Microprocessor (2)
Communications Software
Network Interface Hardware
. . .
Embedded Microprocessor (N)
Communications Software
Network Interface Hardware
High Speed Parallel Switching System (HSPSS)
Large Storage Disk

Fig. 3. High speed parallel processing system architecture.

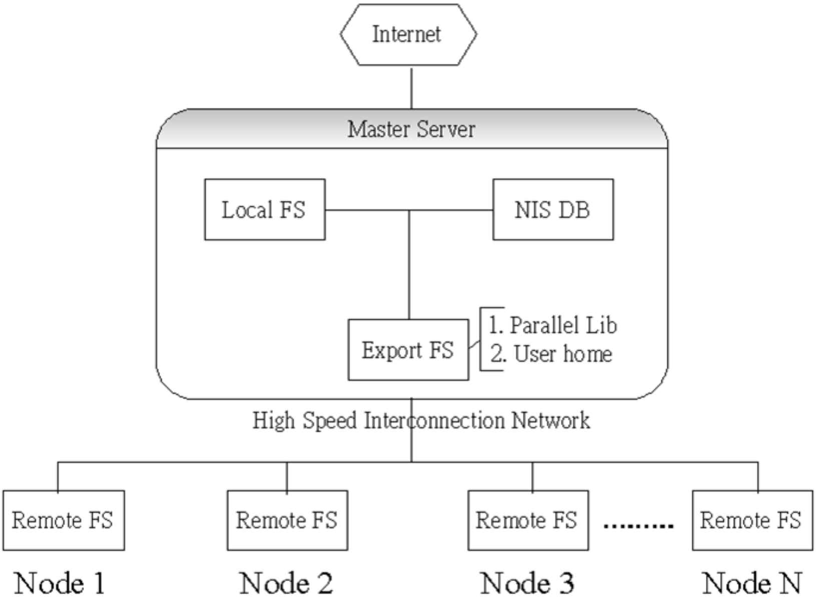


Fig. 4. Software in a cluster computing system.

second (Mbps). The non-integer level numbers are referred as “intermediate levels." All levels have the same status, but note that some applications may choose to use only the integer-numbered levels.

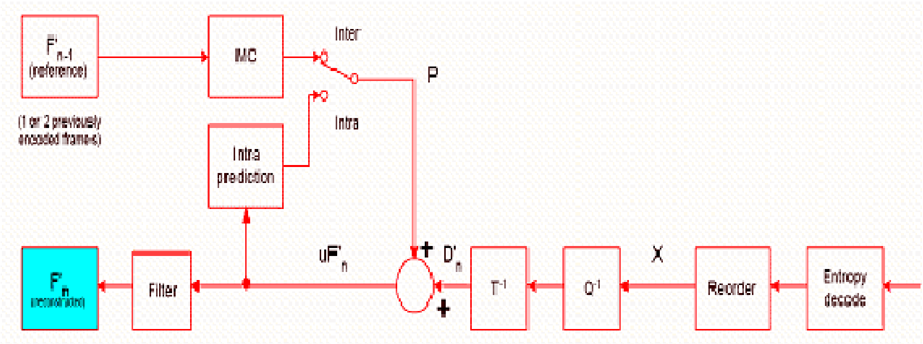


Fig. 5. H.264 decoder architecture.

The H.264 standard codec, JM 8.0, is adopted for use in our proposed wallet-size cluster system. Figure 5 shows the H.264 decoder scheme. The basic functional elements (prediction, transform, quantization, entropy encoding) are slightly different from previous standards (MPEG1, MPEG2, MPEG4, H.261 and H.263); the important changes in H.264 occur in the details of each functional element. The decoder receives a compressed bit stream from the `NAL`. The data elements are entropy decoded and reordered to produce a set of quantized coefficients `X`. These are re-scaled and inverse transformed to give `D’n` (this identical to the `D’n` shown in the Encoder).

Using the header information decoded from the bit stream, the decoder creates a prediction macro block `P`, identical to the original prediction `P` formed in the encoder. `P` is added to `D’n` to produce `uF’n`, that is filtered to create the decoded macro block `F’n`. According to the scheme described, the processing complexity of each module in the decoder can further be estimated. By using the `Intel VTune Performance Analyzer` to analyze the percentage of execution time of each module in the decoder, when the baseline profile is considered.

The result shown in Figure 6 is obtained by decoding a 4CIF resolution video with Qstep 30. The most critical time modules are in sequence motion compensation (41.88%), followed by entropy coding (28.25%), then deblocking (10.48%), and integer transform (10.33%).

Obviously, by applying the proposed scheme using parallel and distributed processing techniques can speedup the video decoding for fulfilling the real time requirement. For the purpose of evaluating the performance of such system, H.264 standard codec JM 8.0 is adopted for our experiment to measure the following data: PSNR, bit rate, complexity reduction of the testing sequences, and decoding time. `Intel Vtune Performance Analyzer` was used to measure the computational cost in terms of number of clock cycles used in each module.

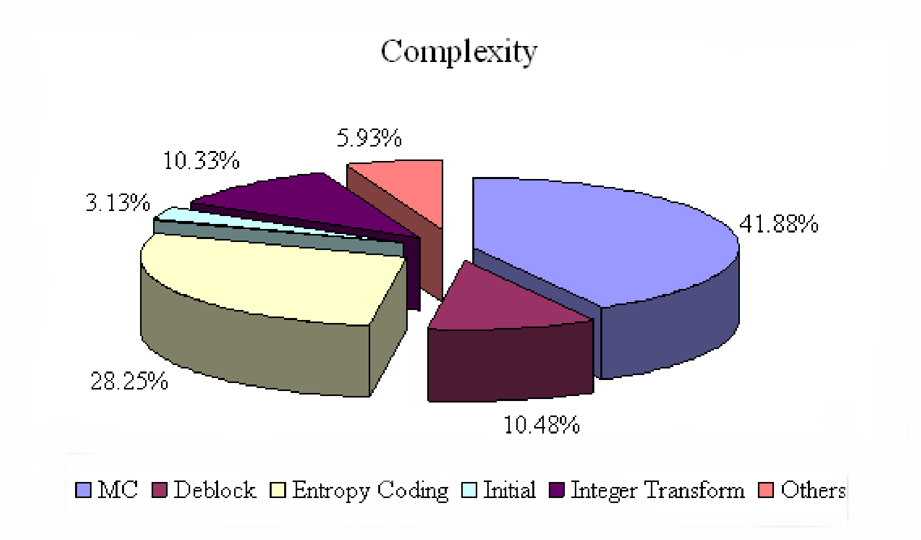


Fig. 6. The complexity of each module in H.264 decoder.

5 Conclusion

Comparing with the previous standards, H.264 is more flexible and possible to improve coding efficiency. However, it should be noted that, this is the expense of added complexity to the encoder/decoder. Moreover, it is not backward compatible to the previous standards.

The level of complexity of the decoder can be reduced by designing the specific profile and level. In this paper, we proposed a scalable wallet-size cluster computing system with the low cost, small die-area and low power consumption. By clustering the embedded system to increase the computing power, the system can be applied to multimedia applications to speedup the video decoding for fulfilling the real time requirements. As future research, we will explore H.264 decoder and the methodology for supporting multimedia communication and buffer management between server and clients. Advances in evaluating the performance of the overall system will also be performed.

Also as future work, a project to integrate available wallet-size cluster system in the campus, in a city or on the internet to form a computing farm by using grid technology, so as way to fully utilize the available and idle systems. Basically, the idea is to develop a grid computing's resource broker, to assist users to identify suitable resources on this specific grid for video conversion or compression, managing efficiently resource usage.

References

1. “ARM7TDMI Data Sheet", http://www.arm.com/products/CPUs/ARM7TDMI.html
2. Chih-Hung Li, Han Lin, Chung-Neng Wang, and Tihao Chiang, “A Fast H.264-Based Picture-In-Picture (PIP) Transcoder", in Proceedings of theICME’2004 - IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, Taiwan, 2004
3. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC), 2003
4. A. Chang, O. C. An, and Y. M. Yeung, “A novel approach to fast multi-frame selection for H.264 video coding", Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, pp. 413-416, 2003
5. Intel Performance Tool VTune homepage, http://www.intel.com/software/products/vtune/
6. K.C. Li, J-L. Gaudiot, and L.M. Sato, “Performance Prediction Methodology for Parallel Programs with MPI in NOW environments", in Proceedings of IWDC’2002 - International Workshop on Distributed Computing, LNCS2571, Springer-Verlag Heidelberg, Sajal K. Das, Swapan Bhattacharya (Eds.), India, 2002
7. http://micetek.com/
8. I.E.G. Richardson, H.264 and MPEG-4 Video Compression, John Wiley & Sons, 2003
9. T. Sterling, G. Bell, and J. S. Kowalik, “Beowulf Cluster Computing with Linux", MIT Press, March 2002
10. A. Tamhankar and K. R. Rao, “An overview of H.264/MPEG-4 Part 10," Proceedings of the 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications, Vol. 1, pp.1-51, 2003
11. C.T. Yang and K.T. Wang, “A Video-on-Demand (VOD) System on Linux High-Availability and Load Balancing Servers", in Proceedings of the ICME’2004 - IEEE International Conference on Multimedia and Expo, Taiwan, 2004

An Adaptive LS-Based Motion Prediction Algorithm for Video Coding

Min-Cheol Hong[1], Myung-Sik Yoo[1], and Ji-Hee Kim[2]

[1] School of Electronic Engineering, Soongsil University, Korea
{mhong,myoo}@e.ssu.ac.kr
[2] Mcube Works, Seoul, Korea
jihee@mcubeworks.com

Abstract. In this paper, we introduce an adaptive motion vector prediction algorithm to improve the performance of a video encoder. The block-based motion vector can be characterized by the local statistics so that the coefficients of LS-based linear motion predictor can be optimized. However, it requires very expensive computational cost, which is major bottleneck in real-time implementation. In order to resolve the problem, we propose the LS-based motion prediction algorithm using spatially varying motion-directed property, so that the coefficients of the motion predictor can be adaptively controlled, resulting in the reduction of computational cost as well as the prediction error. Experimental results show the capability of the proposed algorithm.

1 Introduction

Motion vector predictor has been used to improve coding efficiency in block-based video coding by making the probability of difference between the motion vector and the motion predicted value less variance, so that the transmitted bits for motion vector difference become smaller [1].

Block-based video coding mechanism including existing video coding standards uses median filter as the motion predictor due to its simplicity and efficiency [1]-[2]. The median filter having the motion-directed property is a context-based non-linear prediction, and it has the capability to keep the motion directed property of neighboring blocks. Also, the median predictor has an advantage on that the motion vector is reconstructed by the transmitted differential motion vector and the decoded motion vectors of the neighboring blocks. However, the median motion predictor uses the limited local information, and therefore it may lead to large prediction error and variance when the local statistics of neighboring motion vectors is non-stationary. In such case, the median motion vector predictor may fail to obtain the optimal differential motion vector, resulting in the decline of coding efficiency [3]-[5].

The LS (Least Squares)-based prediction produces locally optimal coefficients by reflecting the spatially varying statistics of motion vector. However, the block-based LS motion prediction requires compute-intensive operations, and therefore direct use of LS-based motion prediction may be bottleneck in the applications for the real-time video delivery.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 705–712, 2004.

In this paper, we propose an adaptive motion vector prediction algorithm to reduce the computational cost with the reduction of the motion prediction error by reinterpreting LS-based prediction from the viewpoint of motion-directed property. The coefficients of the motion predictor are adaptively determined by incorporation of the spatially varying statistics of motion vectors.

This paper is organized as follows. In Section 2, the LS-based motion predictor is formulated. Section 3 explains the motion-directed property of LS-based prediction, and the new adaptive motion vector prediction is introduced. Finally, the experimental results and conclusions are followed in Sections 4 and 5.

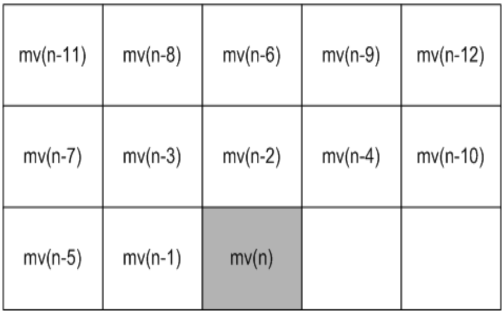
2 Problem Formulation

It is well known that the correlation of motion vectors between neighboring blocks in video coding is very high, and therefore the difference between the real motion vector and the predicted value is encoded and transmitted to improve the coding efficiency. MVD (Motion Vector Difference) can be written as

$$MVD(m,n) = mv(m,n) - \hat{mv}(m,n), \tag{1}$$

where (m,n) represents the block location, and mv and $\hat{mv}$ denote the real motion vector and the predicted motion vector, respectively.

The purpose of motion prediction in video coding is to obtain a motion predicted value which is as close as possible to the real motion vector so that a certain symbol of the difference between the real motion vector and the predicted one has very high probability.



mv(n-11)	mv(n-8)	mv(n-6)	mv(n-9)	mv(n-12)
mv(n-7)	mv(n-3)	mv(n-2)	mv(n-4)	mv(n-10)
mv(n-5)	mv(n-1)	mv(n)		

Fig. 1. Order of motion vectors (N=12)

In our algorithm, it is assumed that the motion vector is stored in stack order as shown in Figure 1, where the n-th motion vector $mv(n)$ is consisted of the horizontal and the vertical components. Each motion vector component is independently predicted in our algorithm. Motion vector prediction using LS based linear prediction considers the N nearest causal motion vector neighbors by the N-th Markovian property in clockwise order. Then, one-step-ahead linear prediction can be written as

$$\hat{mv}(n) = \sum_{k=1}^{N} a(k)mv(n-k), \tag{2}$$

where $\hat{mv}(n)$ and $mv(n-k)$ represent the predicted motion vector and the k-th prediction neighbor of the n-th block, respectively. Also, the weights $a(k)$ denotes the k-th linear predictor coefficients.

In general, motion vectors in moving object regions are characterized by abrupt changes of local statistics, which violates stationary assumption to optimize the coefficients of LS-based linear motion prediction. Therefore, an adaptive prediction approach is motivated to handle such non-stationary nature of source. More specifically, it is reasonable to estimate the local statistics such as classical covariance method under the assumption that the source is locally stationary. Suppose that the causal window called as "training window" is used to estimate the covariance, and it contains $M = 2T(T+1)$ causal motion vector neighbors as shown in Figure 2.

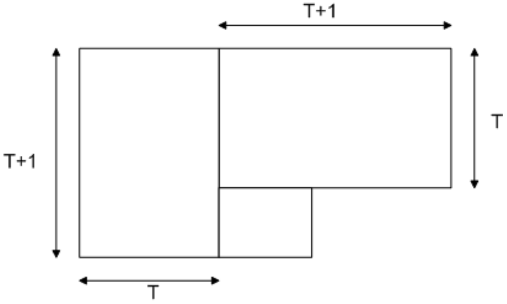


Fig. 2. Training window

Consider the training window by a $M \times 1$ column vector $\boldsymbol{mv} = [mv(n-1), mv(n-2), \ldots, mv(n-M)]^T$, where T represents transpose operation. Then, $\mathbf{C}$, the motion vector prediction neighbors of $\boldsymbol{mv}$, takes the form of an $M \times N$ matrix

$$\mathbf{C} = \begin{bmatrix} mv(n-1-1) & mv(n-1-2) & \ldots & mv(n-1-N) \\ mv(n-2-1) & mv(n-2-2) & \ldots & mv(n-2-N) \\ . & . & \ldots & . \\ mv(n-M-1) & mv(n-M-2) & \ldots & mv(n-M-N) \end{bmatrix}, \tag{3}$$

where $mv(n-j-k)$ is the k-th prediction neighbor of $mv(n-j)$. Then, the motion prediction coefficients can be obtained by LS optimization inside the training window. It is [6]-[10]

$$min||\boldsymbol{mv} - \mathbf{C}\boldsymbol{a}||^{\mathbf{2}}, \tag{4}$$

where $||\cdot||$ represents the Euclidean norm, and $\boldsymbol{a} = [a(1), a(2), \ldots, a(N)]^T$ denotes the optimized motion prediction coefficients. Then, the LS optimization has the following closed form solution.

$$\boldsymbol{a} = (\mathbf{C^T C})^{-1}(\mathbf{C^T}\boldsymbol{mv}). \tag{5}$$

3 Proposed LS-Based Motion Prediction Algorithm

The effectiveness of an adaptive motion vector prediction scheme depends on its capability of adapting from motion smooth region to motion edge areas. The difficulty of achieving ideal adaptation mainly arises from the motion edge areas because the orientation of a motion edge can be arbitrary.

Motion vectors of neighboring blocks within the training window can be classified into the motion edge neighbors around the moving areas and the motion non-edge neighbors away from the moving areas. For the motion non-edge neighbors, the matrix $\mathbf{C}$ is often not full-ranked and the LS optimization does not have a unique solution. In fact, the set of optimal motion predictors for the motion non-edge neighbors lies in a hyper-plane $\sum_{k=1}^{N} a(k)$ in the N-dimensional space. On the other hands, the matrix $\mathbf{C}$ is usually full-ranked and the LS optimization has a unique solution for the motion edge neighbors. It is of easy to see that the set of optimal motion predictors for the motion edge neighbors is a subset of the hyper-plane $\sum_{k=1}^{N} a(k)$. Consequently, the motion edge neighbors dominate the LS optimization process. The LS optimization over the training window offers a convenient way to obtain the optimal prediction coefficients for the motion edge neighbors without the necessity of motion edge detection [7].

The bottleneck of the LS optimization is compute-intensive operations of the covariance matrix $\mathbf{C^T C}$ in Eq. (5). In this paper, we propose an adaptive motion vector prediction to reduce the computational cost without the loss of coding efficiency by performing the LS optimization only for a fraction of the motion vectors in the video frames. The motion vector prediction using the motion edge-directed property is based on the following two observations. First, the motion prediction coefficients optimized for a motion vector around a motion edge are often suitable for its neighbors along the same motion edge. Second, the set of optimal motion predictors for a motion edge is the subset of the set of optimal motion predictors for the motion smooth region. Therefore, the motion prediction coefficients optimized for a motion edge can be stored and repeatedly used until the next motion edge. In other words, LS optimization performs on a motion edge-by-motion edge basis rather than on a motion vector-by-motion vector basis. In this paper, to implement the LS optimization on motion edge basis, the following switching strategy is defined.

$$e(n) = mv(n) - \hat{mv}(n-1). \tag{6}$$

If the motion prediction error is beyond a pre-selected threshold, the LS optimization is activated to update the motion prediction coefficients such as

$$a(n) = \begin{cases} ((\mathbf{C^T C})^{-1}(\mathbf{C^T}\boldsymbol{mv}))(\mathbf{n}) & \text{if } |e(n)| > TH \\ a(n-1) & \text{otherwise} \end{cases}, \tag{7}$$

where $((\mathbf{C^T C})^{-1}(\mathbf{C^T}\boldsymbol{mv}))(n)$ represents the n-th component of $(\mathbf{C^T C})^{-1}(\mathbf{C^T}\boldsymbol{mv})$. Then, the motion predicted value, $\hat{mv}(n)$ in Eq. (2) is updated as

$$\hat{mv}(n) = \begin{cases} \sum_{k=1}^{N} a(k)mv(n-k) & \text{if } |e(n)| > TH \\ \hat{mv}(n-1) & \text{otherwise} \end{cases}. \tag{8}$$

4 Experimental Results

JM2 (Joint Model Number 2) of H.264 video coding standard was used in order to compare the performance of the proposed algorithm with the median filter. The proposed algorithm was tested with various video sequences and resolution at a number bit rates. In our experiments, the four adjacent neighbors in the training window are used and the threshold is 20. For evaluating the performance of the algorithm, PSNR (Peak Signal to Noise Ratio) is used. For $M \times N$ dimensional 8-bit image, it is defined as

$$PSNR = 10log\frac{MN \times 255^2}{||f-\tilde{f}||^2}, \tag{9}$$

where f and $\tilde{f}$ denote the original image and the reconstructed image, respectively, and $||.||$ represents the Euclidean norm. In addition, MPEPB (Motion Prediction Error Per Block) is defined to evaluate the accuracy of the motion predictor. It is written as

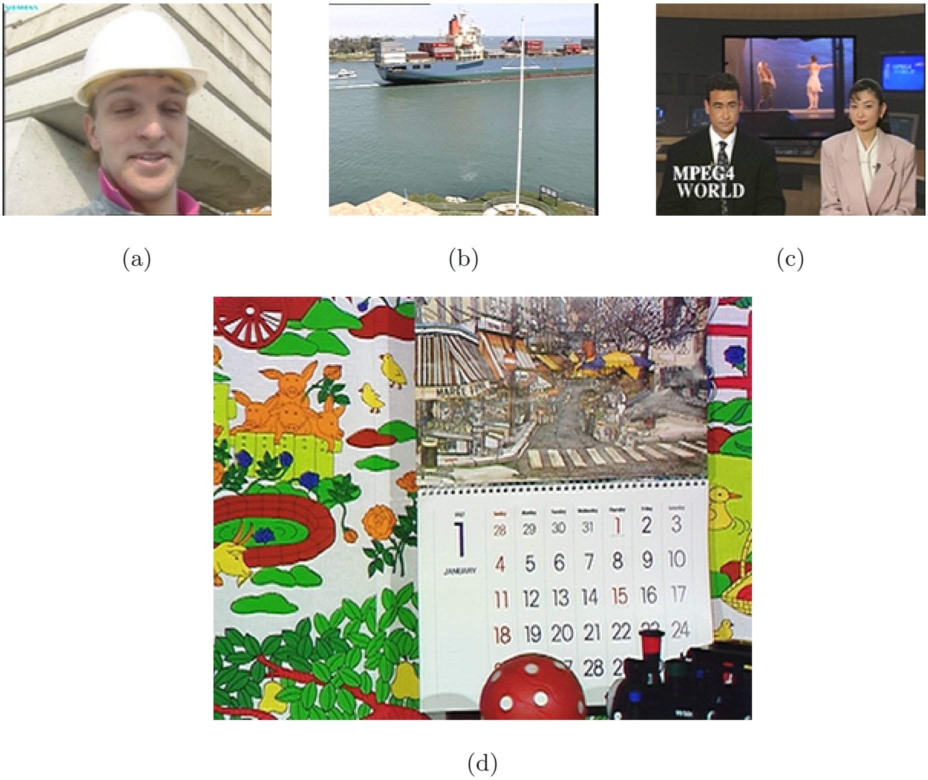
$$MPEPB = \frac{1}{K}\sum_{i=1}^{K}(|mv_x(i) - \hat{mv}_x(i)| + |mv_y(i) - \hat{mv}_y(i)|), \tag{10}$$

where K is the number of total blocks in an image. Also, $mv_j(i)$ and $\hat{mv}_j(i)$ denote the estimated and the predicted motion values of i-th block in horizontal and vertical directions, respectively. Also, the evaluating the computational complexity, ETSMP (Encoding Time Saving of Motion Prediction) is defined as

$$ETSMP = \frac{ET(A) - ET(B)}{ET(A)} \times 100(\%). \tag{11}$$

where $ET(A)$ and $ET(B)$ represent the encoding times with median filter and the proposed algorithm. The plat-form used is Pentium III 700 MHz with 256 MB RAM. Full search motion estimation was used to determine the block-based motion vector, and the motion search range to the horizontal and the vertical directions was 32. In the set of the experiments, QCIF Foreman sequence with 10 frames/sec, QCIF Container sequence with 10 frames/sec, QCIF News sequence with 10 frames/sec, and CIF Mobile sequence with 30 frames/sec are described, which are shown in Figure 3.

Table 1 shows the bit rates and the PSNR comparisons as a function of Quantization index. The results illustrate that the marginal PSNR gains are consistently obtained against the median filter, since as shown in Table 2 the



(a) (b) (c)

(d)

Fig. 3. Test sequences used for experiments, (a) QCIF Foreman sequence, (b) QCIF Container sequence, (c) QCIF News sequence, (d) CIF Mobile sequence

motion prediction error of the proposed algorithm is smaller than the median filter so that the amount of transmitted bits for motion information is reduced. Also, Table 2 describes that with the proposed algorithm the complexity gain is reduced, and that the complexity gain is higher as the motion between blocks is more stationary.

The novelty of the proposed algorithm is in that the motion vector prediction coefficients are adaptively controlled by motion edge-directed property of its neighbors, resulting in the subset of the set of optimal motion predictors for the motion smooth region.

5 Conclusions

In this paper, we presented an adaptive motion vector prediction algorithm using the LS optimization in motion edge neighbors. The coefficients of the motion prediction are adaptively determined by using spatially varying motion-directed property. It results in reducing the computational complexity with the marginal PSNR improvement.

Table 1. Bit rates and PSNR comparisons as a function of Quantization index

Sequence	method	QP12		QP16		QP20		QP24	
		bit rates (kbps)	PSNR (dB)	bit rates (kbps)	PSNR (dB)	bit rates (kbps)	PSNR (dB)	bit rates (kbps)	PSNR (dB)
Foreman (QCIF, 10fps)	median filter	79.21	36.04	48.85	33.48	30.30	30.96	18.54	28.26
	proposed method	79.20	36.07	48.83	33.52	30.29	30.99	18.54	28.28
News (QCIF, 10fps)	median filter	49.31	36.83	30.34	33.87	18.27	31.02	10.49	28.16
	proposed method	49.27	36.88	30.33	33.91	18.25	31.05	10.46	28.16
Container (QCIF, 10fps)	median filter	26.80	36.08	14.34	33.50	8.27	30.70	4.73	27.95
	proposed method	26.77	36.11	14.33	33.53	8.25	30.72	4.72	27.95
Mobile (CIF, 30fps)	median filter	1974.3	33.89	1025.6	30.61	491.5	27.51	247.9	24.52
	proposed method	1969.8	33.91	1021.4	30.64	488.9	27.53	245.8	24.54

Table 2. MPEPB and ETSMP comparisons as a function of Quantization index

Sequence	method	QP12		QP16		QP20		QP24	
		MPE PB	ETS MP	MPE PB	ETS MP	MPE PB	ETS MP	MPE PB	ETS MP
Foreman (QCIF, 10fps)	median filter	3.1	N/A	3.2	N/A	3.2	N/A	3.3	N/A
	proposed method	2.5	19.35	2.5	19.41	2.6	19.44	2.7	19.08
News (QCIF, 10fps)	median filter	2.6	N/A	2.6	N/A	2.7	N/A	2.7	N/A
	proposed method	1.9	20.63	1.9	20.96	1.9	19.67	2.1	21.31
Container (QCIF, 10fps)	median filter	2.1	N/A	2.3	N/A	2.3	N/A	2.5	N/A
	proposed method	1.4	19.35	1.4	19.67	1.5	21.31	1.6	20.08
Mobile (QCIF, 10fps)	median filter	4.6	N/A	4.5	N/A	4.6	N/A	4.7	N/A
	proposed method	3.8	21.51	3.9	21.26	4.0	21.12	4.0	21.57

Approaches to adjust the number of adjacent motion vector neighbors in training window and to adaptively determine threshold using the given information of video images are under investigation. With incorporation of the information, it is expected that better results can be obtained.

Acknowledgements. This work was supported by grant No. R01-2002-000-00073-0 from the Basic Research Program of the Korean Science and Engineering Foundation.

References

1. K. Saywood, *Introduction to Data Compression,* Morgan Kaufmann, 2000
2. ITU-T and ISO/IEC, "Final Committee Draft (CD) of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14996-10 AVC)," Dec. 2002
3. S. Chakravarti, T. P. Jung, S. C. Ahalt, and A. K. Krishnamurthy, "Comparison of prediction methods for differential image processing application," Proceeding of International Conference on System Engineering, pp. 210–213, 1991
4. S. D. Kim and J. B. Ra, "An efficient motion vector coding scheme based on minimum bit rate prediction," IEEE Trans. on Image Processing, vol. 8, pp. 1117–1120, Aug. 1999
5. D. H. Kang, J. H. Choi, Y. H. Lee, and C. Lee, "Application of a DPCM system with median predictors for image coding," IEEE Trans. on Consumer Electronics, vol. 38, pp. 429–435, Aug. 1992
6. X. Wu and N. Memon, "Context-based adaptive lossless image coding," IEEE Trans. on Communication, vol. 45, pp. 437–444, Apr. 1997
7. M. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," IEEE Trans. on Image Processing, vol. 9, pp. 1309–1324, Aug. 2000
8. X. Li and M. Orchard, "Edge-directed prediction for lossless compression of natural images," IEEE Proceeding of International Conference on Image Processing, vol. 4, pp. 58–62, Oct. 1999
9. N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Englewood Cliffs, NJ: Prentice-Hall, 1984
10. H. Ye, G. Deng, and J. Devlin, "Least squares approach for lossless image coding," Proceeding of Signal Processing Applications, vol. 1, pp. 63–66, 1999

Embedded Packetization Framework for Layered Multiple Description Coding

Longshe Huo[1,2], Qingming Huang[2], and Jianguo Xie[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] Graduate School, Chinese Academy of Sciences, Beijing, China
{lshuo,qmhuang,jgxie}@jdl.ac.cn
http://www.jdl.ac.cn

Abstract. We present an embedded packetization framework for layered multiple description codes, in which both the base layer and the enhancement layer share the same number of packets, while each packet is partitioned into two parts, each belongs to one layer. In this framework, optimizing a single multiple description code for high-bandwidth clients and directly applying it to low-bandwidth clients only result in slight performance degradation for the latter. We also propose two local search algorithms, with one extending the solution optimized for low-bandwidth clients to be available for high-bandwidth clients, and the other improving the weighted average performance of both high- and low-bandwidth clients. Our analysis shows that the performance gains are significant, and a better performance tradeoff among all clients can be achieved than previous solutions.

1 Introduction

Multiple description coding (MDC) [1] has recently emerged as an attractive framework for robust multimedia transmission over unreliable channels. Many methods of MDC have been developed over the years. One particularly efficient and practical method is based on priority encoding transmission (PET) technique [2], which combines scalable source coding with unequal error protection (UEP) to minimize the impact of lost packets on the quality of network service. The idea is to partition a scalable source bitstream into segments of decreasing importance, and protect these segments using progressively weaker forward error correction (FEC) channel codes, so as to convert a scalable, prioritized bitstream into multiple non-prioritized descriptions (packets), and achieve the best joint economy of source and channel codes. We call this MDC method as FEC-MDC in this paper.

Chou et al. [3] proposed codes which split multiple descriptions of FEC-MDC into layers, and introduced the concept of layered multiple description coding (LMDC). This technique has the advantages of both layered codes and multiple description codes, as it allows low bandwidth clients to receive a base MDC layer while high bandwidth clients to receive both a base and an enhancement MDC layer. In their construction, packets are partitioned into two parts, such

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 713–720, 2004.

that different layers consist of different number of packets, while each layer has the same packet length. In this scenario, optimizing a single FEC-MDC for one layer and directly applying it to another layer may result in a potentially large distortion for the latter. In [3], Chou et al. first optimized the base layer packets for low-bandwidth clients, then proposed two methods to optimize the additional packets in the enhancement layer to minimize the distortion for high-bandwidth clients. Even using these two methods, they declared that the best overall performance of high-bandwidth clients was still 1.4 dB away from the minimum possible distortion optimized under non-layered environment.

In this paper, we present a new packetization framework for LMDC. We partition the layers across vertical direction, i.e., every packet is split into two parts belonging to base layer and enhancement layer respectively. In this scheme, optimizing a single FEC-MDC for high-bandwidth clients and directly applying it to low-bandwidth clients only result in slight performance degradation for the latter. We also propose two local search algorithms, with one extending the solution optimized for low-bandwidth clients to be available for high-bandwidth clients, and the other improving the weighted average performance of both high- and low-bandwidth clients.

In Section 2, we review the background of FEC-MDC and LMDC. In Section 3, we present our new LMDC packetization framework and two local search algorithms. In Sections 4, we present our results, and in Section 5 we present our conclusions.

2 Background

2.1 FEC Based Multiple Description Coding

For scalable video coding, the original video sequence is often partitioned into groups of frames (GOF), with each GOF containing a fixed number of frames and being encoded into an independent embedded bitstream. Consider transmission of an embedded source bitstream over a packet loss channel using N packets of L bytes each. In the FEC-MDC framework, the source bitstream is divided into L consecutive segments $S_1, ..., S_L$ of $m_i \in \{1, ..., N\}$ bytes each and each segment is protected by an (N, m_i) systematic RS code. Let $f_i = N - m_i$ denote the number of RS redundancy bytes that protect segment S_i, $1 \leq i \leq L$. If n packets of N are lost, the RS codes ensure that all segments that contain at most $N - n$ source bytes can be recovered. Since the embedded source bitstream is sequentially refinable, decoding of the S_i segment depends on all the previous $i - 1$ segments, thus the number of redundancy bytes must be monotonically non-increasing in the segment index, i.e., $f_1 \geq ... \geq f_L$. Under this constraint, if at most f_i packets are lost, then the receiver can decode at least the first i segments. In this paper, we use an L-dimensional vector $F_L = (f_1, ..., f_L)$ to denote a FEC-MDC protection scheme (FPS), where $f_i \in \{0, ..., N-1\}$ and $f_1 \geq ... \geq f_L$. Let $p_N(n)$ denote the probability of losing exactly n packets out of N and let $c_N(k) = \sum_{n=0}^{k} p_N(n)$, $k = 0, ..., N$, then $c_N(f_i)$ is the probability that the receiver correctly recovers segment S_i. Let $\phi(r)$ be the rate-distortion function of the embedded source bitstream, which is a monotonically non-increasing function, then the expected distortion of the reconstructed sequence at the decoder side can be expressed as

$$E_D(F_L) = c_N(N)\phi(0) - \sum_{i=1}^{L} c_N(f_i)(\phi(r_{i-1}) - \phi(r_i)), \tag{1}$$

where $r_i = \sum_{k=1}^{i} m_k = iN - \sum_{k=1}^{i} f_k, 1 \leq i \leq L$. Hence the objective of this problem is to find the FPS $F_L = (f_1, ..., f_L)$ that minimizes (1), for given N, L, $p_N(n)$, and $\phi(r)$.

Several researchers devised efficient algorithms addressing this problem [4, 5, 6, 7, 8]. Dumitrescu et al. [7] presented a globally optimal algorithm, but its complexity was too high in both time and space to be used in real time. In all experiments of this paper, we use the algorithm of Mohr et al. [5], whose performance is near optimal and time complexity is also acceptable.

2.2 Layered Multiple Description Coding

Chou et al. [3] partitioned FEC-MDC into layers and constructed layered multiple description codes. In their constructions, the base MDC layer consists of N_1 packets per GOF, while the enhancement MDC layer consists of N_2 packets per GOF, thus that each packet has a fixed length of L bytes. The base MDC layer is transmitted to each low-bandwidth client, while both the base and enhancement MDC layers are transmitted to each high-bandwidth client. The first N_1 packets are shared by both low-bandwidth and high-bandwidth codes. We call this framework as Fixed-length Packetization Framework (FPF).

There are two naive methods to construct an LMDC in this framework. One is to optimize a single FEC-MDC for high-bandwidth clients, and split it into base and enhancement layers by transmitting only the first N_1 packets to low-bandwidth clients. This may result in a large distortion for low-bandwidth clients. The other method is to optimize a single FEC-MDC for low-bandwidth clients, and transmit all of it, plus N_2 additional parity packets, to high-bandwidth clients. This may give a large distortion for high-bandwidth clients. For convenience, following we denote these two methods as FPF-A and FPF-B respectively.

In [3], Chou et al. presented two additional methods to alleviate the performance degradations. Both methods first optimize the N_1 packets of the MDC base layer for low-bandwidth clients, and then optimize other N_2 packets in the MDC enhancement layer to minimize the distortion for high-bandwidth clients, using two different constructions. The idea of the first one, following called FPF-C, is to borrow some number q of the N_2 packets from the enhancement layer to protect the base layer as additional parity packets, while the other $N_2 - q$ packets in the enhancement layer are used to protect the remaining source bytes not already present in the base layer. The second method repeats some of the less protected bytes from the end of the base layer in the enhancement layer, protects them together with the remaining source bytes not already present in the base layer using all N_2 packets. Both these two methods can achieve the possible best performance for low-bandwidth clients, but still 1.4dB worse than their minimum possible distortion for high-bandwidth clients, declared by Chou et al. in [3], where the source coder used was MPEG4-FGS. The results are unacceptable for situations in which the population of high-bandwidth clients is greater than that of low-bandwidth clients.

3 Embedded Packetization Framework for LMDC

We construct LMDC using a different way from above. In our construction, both the base layer and the enhancement layer consist of the same N packets per GOF, while each packet is partitioned into two parts, with the first L_1 bytes belonging to the base layer and the remaining L_2 bytes belonging to the enhancement layer. All bytes of each packet are transmitted to high-bandwidth clients, while only the first L_1 bytes of each packet are first truncated and grouped into a new packet, and then transmitted to low-bandwidth clients. In this construction, each packet of low-bandwidth MDC codes are embedded in the packets of high-bandwidth MDC codes, thus we denote it as Embedded Packetization Framework (EPF).

In this framework, for a given LMDC construction, the FPS for low-bandwidth clients must be a prefix of the FPS for high-bandwidth clients. Thus the problem of constructing an LMDC becomes to find only an FPS for this LMDC that optimizes for both high-bandwidth clients and low-bandwidth clients jointly.

For convenience, we denote the FPS solely optimized for low-bandwidth clients as $F^B_{L_1} = (f^B_1, ..., f^B_{L_1})$, and the FPS solely optimized for high-bandwidth clients as $F^H_{L_1+L_2} = (f^H_1, ..., f^H_{L_1}, f^H_{L_1+1}, ..., f^H_{L_1+L_2})$. Now there are three methods we can use to construct a LMDC. The first one, denoted as EPF-A, is to directly use $F^H_{L_1+L_2}$ as the FPS of this LMDC. In this case, high-bandwidth clients can reach their possible minimum distortion, while low-bandwidth clients may suffer slight performance losses. The second way is to use $F^B_{L_1}$ as a prefix of this LMDC's FPS, and then extends it to the length of $L_1 + L_2$. This may result in the possible minimum distortion for low-bandwidth clients, and also slight performance degradations for high-bandwidth clients. The final method is to construct a FPS optimized neither solely for high-bandwidth clients nor solely for low-bandwidth clients, but is optimal for the weighted average performance of them. In this section, we present two local search algorithms for the latter two methods.

3.1 Local Search Algorithm by Extending the FPS Optimized for Base Layer

Assume the first L_1 elements of an LMDC FPS $F_{L_1+L_2}$ have been determined by $F^B_{L_1}$, now we consider the problem of finding other L_2 elements to minimize the expected distortion seen by high-bandwidth clients. Our algorithm is inspired from the local search method of Stankovic et al. [8]. Let $F_{L_1+L_2} = (f_1^B, ..., f_{L_1}^B, f_{L_1+1}, ..., f_{L_1+L_2})$, there must be $f_{L_1}^B \geq f_{L_1+1} \geq ... \geq f_{L_1+L_2}$. Thus we can start from the solution with $f_{L_1+1} = ... = f_{L_1+L_2} = f_{L_1}^B$ and iteratively decrease the protection strength to find the possible best solution.

Definition 1. *Let $F_{L_1+L_2} = (f_1^B, ..., f_{L_1}^B, f_{L_1+1}, ..., f_{L_1+L_2})$ be a feasible solution for a given LMDC. The neighborhood of $F_{L_1+L_2}$ consists of the solutions of the form: $(f_1^B, ..., f_{L_1}^B, f_{L_1+1}, ..., f_{L_1+L_2-1}, f_{L_1+L_2} - 1), (f_1^B, ..., f_{L_1}^B, f_{L_1+1}, ..., f_{L_1+L_2-1} - 1, f_{L_1+L_2} - 1), ..., (f_1^B, ..., f_{L_1}^B, f_{L_1+1} - 1, ..., f_{L_1+L_2-1} - 1, f_{L_1+L_2} - 1)$, which are also feasible solutions for the LMDC.*

Based on this definition, we give a local search algorithm as follows.

Algorithm 1 (EPF-B). Local search by extending the FPS optimized for base layer.

1. Initializes current feasible solution: $F_{L_1+L_2} = (f_1^B, ..., f_{L_1}^B, f_{L_1+1} = f_{L_1}^B, ..., f_{L_1+L_2} = f_{L_1}^B)$.
2. If $f_{L_1+L_2} = 0$, stop and return $F_{L_1+L_2}$ as the best solution.
3. Search for the solution $F'_{L_1+L_2}$, whose expected distortion is the minimum in the neighborhood of $F_{L_1+L_2}$.
4. If $E_D(F'_{L_1+L_2}) < E_D(F_{L_1+L_2})$, set $F_{L_1+L_2} = F'_{L_1+L_2}$ and go to Step 2; else stop and return $F_{L_1+L_2}$ as the best solution.

For convenience, we denote the FPS optimized using this algorithm as $F_{L_1+L_2}^{B'} = (f_1^B, ..., f_{L_1}^B, f_{L_1+1}^{B'}, ..., f_{L_1+L_2}^{B'})$.

3.2 Local Search Algorithm by Optimizing Weighted Average Performance

If an LMDC is optimized for high-bandwidth clients, it is not fair for low-bandwidth clients especially when their population is greater than that of the former, and vice verse. In this section, we define a weighted average performance measurement, and propose a local search algorithm to optimize it. Assume the fraction of high-bandwidth clients in total population is h, and the fraction of low-bandwidth clients is $1 - h, 0 \leq h \leq 1$, then for a given LMDC, its weighted average expected distortion is defined as

$$WE_D(F_{L_1+L_2}, h) = h \cdot E_D(F_{L_1+L_2}) + (1 - h) \cdot E_D(F_{L_1}), \qquad (2)$$

where F_{L_1} is a prefix of $F_{L_1+L_2}$.

The solutions obtained from the first two methods, $F_{L_1+L_2}^H$ and $F_{L_1+L_2}^{B'}$, can be seen as two special cases by minimizing (2) with $h = 1$ and $h = 0$ respectively. Experiments show that the protection strength of $F_{L_1+L_2}^H$ is stronger than that of $F_{L_1+L_2}^{B'}$. We also proofed this observation in [9], assume the rate-distortion function is strictly convex. If we denote the optimal FPS of minimizing (2) in general case as $F_{L_1+L_2}^W = (f_1^W, ..., f_{L_1+L_2}^W)$, then we guess that the curve of

$F^{W}_{L_1+L_2}$ is between that of $F^{H}_{L_1+L_2}$ and $F^{B'}_{L_1+L_2}$, i.e., the protection strength of $F^{W}_{L_1+L_2}$ is stronger than that of $F^{B'}_{L_1+L_2}$ and weaker than that of $F^{H}_{L_1+L_2}$. Based on this conjecture, we devise two local search algorithms to resolve $F^{W}_{L_1+L_2}$, one starting from $F^{B'}_{L_1+L_2}$ and iteratively increasing its protection strength, while the other starting from $F^{H}_{L_1+L_2}$ and iteratively decreasing its protection strength. Due to limitations in space, following we only describe the first one.

Define $f_0 = N$ and $f_{L_1+L_2+1} = -1$. For $1 \leq i \leq L_1+L_2+1$, if $f_i \neq f_{i-1}$, we call i a redundancy change point. In each iteration, we search the neighborhood of current feasible solution from every possible redundancy change point.

Definition 2. *Let* $F_{L_1+L_2} = (f_1, ..., f_{L_1+L_2})$ *be a feasible solution for a given LMDC. The neighborhood of* $F_{L_1+L_2}$ *at redundancy change point* i*,* $1 \leq i \leq L_1+L_2$*, consists of the solutions of the form:* $(f_1, ..., f_{i-1}, f_i+1, f_{i+1}, ..., f_{L_1+L_2})$, $(f_1, ..., f_{i-1}, f_i+1, f_{i+1}+1, ..., f_{L_1+L_2}), ..., (f_1, ..., f_{i-1}, f_i+1, f_{i+1}+1, ..., f_{L_1+L_2}+1)$*, which are also feasible solutions for the LMDC.*

Algorithm 2 (EPF-C). Local search by optimizing weighted average performance.

1. Initializes current feasible solution: $F_{L_1+L_2} = F^{B'}_{L_1+L_2}$, set $cont = 1$.

2. If $cont = 0$, stop and return $F_{L_1+L_2}$ as the best solution; else set $cont = 0$, $i = 1$.

3. If $f_i = N - 1$, go to Step 6.

4. Search for the solution $F'_{L_1+L_2} = (f_1, ..., f_{i-1}, f_i+1, ..., f_{i+j}+1, f_{i+j+1}, ..., f_{L_1+L_2})$, whose weighted average expected distortion is the minimum in the neighborhood of $F_{L_1+L_2}$ at redundancy change point i.

5. If $WE_D(F'_{L_1+L_2}, h) < WE_D(F_{L_1+L_2}, h), setF_{L_1+L_2} = F'_{L_1+L_2}, cont = 1$, and $i = i + j + 1$.

6. Repeat $i = i + 1$, until i is a redundancy change point. If $i \leq L_1 + L_2$, go to step 4; else go to Step 2.

4 Results

In our experiments, we use a two-state Markov process to simulate the channel model, with mean packet loss probability being 0.1 and mean burst length being approximately 11. The derivation of $p_N(n)$ for this model can be found in [10]. The operational rate-distortion function $\phi(r)$ was obtained by encoding the first 16 frames of the video sequence Foreman (CIF format) as a GOF, using the 3D SPIHT scalable video codec [11]. We choose $N = 64$, $L_1 = L_2 = 625$ bytes for EPF framework, and $N_1 = N_2 = 32$, $L = 1250$ bytes for FPF framework, which is the same as used in [3].

Figure 1 shows the FPSs resolved by method EPF-A, algorithm EPF-B, and algorithm EPF-C with $h = 0.5$. The curve of EPF-A is above that of EPF-B, which means that the protection strength of the FPS resolved by EPF-A is

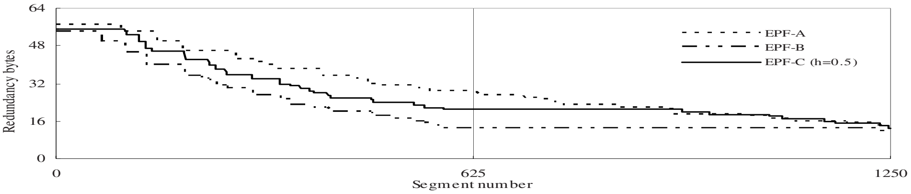


Fig. 1. FEC-MDC protection schemes resolved by EPF-A, EPF-B and EPF-C

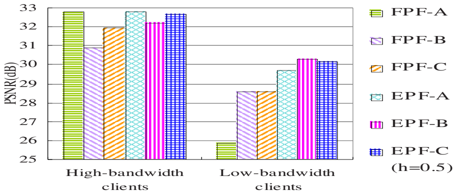


Fig. 2. PSNR of the expected distortion seen by high- and low-bandwidth clients, achieved using various methods

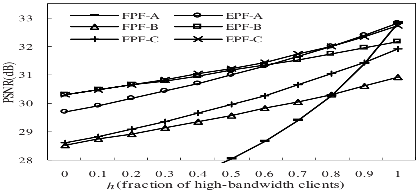


Fig. 3. PSNR of the weighted average distortion, achieved using various methods

stronger than the one resolved by EPF-B. The curve of EPF-C is situated almost in the middle of other two curves. This is accordant with our imagination.

Figure 2 compares the performance of different methods in FPF and EPF frameworks seen by high- and low-bandwidth clients. For high-bandwidth clients, both FPF-A and EPF-A can achieve the possible highest PSNR; FPF-B has a large distortion of nearly 2 dB worse than the best one; FPF-C improves it, but still has a gap of 0.9 dB worse than the best one; the performance of EPF-B is slightly better than that of FPF-C, moreover its complexity is much lower than that of the latter; the performance of EPF-C with $h = 0.5$ is very close to the best one, the difference between their PSNRs is only 0.18 dB. For low-bandwidth clients, the possible best PSNR achieved by FPF-B and FPF-C is less than that of EPF-B, because the parameters they used, N and L, are different; FPF-A has a large distortion of nearly 3 dB worse than that of FPF-B; the PSNRs achieved by EPF-A and EPF-C with $h = 0.5$ are only 0.6 dB and 0.15 dB less than the best one achieved by EPF-B respectively. Though the solution of algorithm EPF-C is optimized neither for high-bandwidth clients nor for low-bandwidth clients, its performances for high- and low-bandwidth clients are very close to their possible best ones, suffering no more than 0.2 dB penalties.

Figure 3 compares the weighted average performance achieved by various methods as parameter h changes from 0 to 1. It can be seen that the weighted average performances of all methods in FPF framework are obviously worse than that in EPF framework. In EPF framework, EPF-A can achieve good average performance when h is near to 1; EPF-B can achieve good average performance when h is near to 0; while EPF-C can achieve good average performance almost for all values of h between 0 and 1, only slightly worse than EPF-A when h comes very near to 1.

5 Conclusion

We have presented an embedded packetization framework for LMDC, in which even naive methods can achieve good performance. We have also proposed two local search algorithms: the first one extends the solution optimized for low-bandwidth clients to be available for high-bandwidth clients, while the second one optimizes the weighted average performance of both high- and low-bandwidth clients. Our analysis shows that the proposed scheme can achieve significant performance gain; especially the second algorithm can achieve near-best performance for both high- and low-bandwidth clients, and fairly adapt to changing population distributions.

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China under grant No. 60333020, the Natural Science Foundation of Beijing under grant No. 4041003, the National Hi-Tech Development 863 Program of China under grant No. 2002AA118010, and the National Hi-Tech Development 863 Program of China under grant No. 2002AA119010. We Thank Dr. Ying Cai for helpful comments.

References

1. Goyal, V.K.: Multiple Description Coding: Compression Meets the Network. IEEE Signal Processing Magazine, 2001, 18(5): 74–93
2. Albanese, A., Blomer, J., Edmonds, J., Luby, M., Sudan, M.: Priority Encoding Transmis-sion. IEEE Trans. on Information Theory, 1996, 42(6): 1737–1744
3. Chou, P.A., Wang, H.J. Padmanabhan, V.N.: Layered Multiple Description Coding. Proc. Packet Video Workshop, Nantes, France, 2003
4. Mohr, A.E., Riskin, E.A., Ladner, R.E.: Graceful Degradation Over Packet Erasure Channels Through Forward Error Correction. Proc. Data Compression Conf., Snowbird, UT, USA, 1999
5. Mohr, A.E., Ladner, R.E., Riskin, E.A.: Approximately Optimal Assignment for Unequal Loss Protection. Proc. IEEE ICIP-2000, Vancouver, Sept. 2000
6. Puri, R., Ramchandran, K.: Multiple Description Coding Using Forward Error Correction Codes. Proc. 33rd Asilomar Conf. on Signals and Systems, Pacific Grove, CA, Oct. 1999
7. Dumitrescu, S., Wu, X., Wang, Z.: Globally Optimal Uneven Error-protected Packetization of Scalable Code Streams. Proc. Data Compression Conf., Snowbird, UT, USA, April 2002
8. Stankovic, V., Hamzaoui, R., Xiong, Z.: Packet Loss Protection of Embedded Data with Fast Local Search. Proc. IEEE ICIP-2002, Rochester, NY, USA, Sep. 2002
9. Huo, L., Gao, W., Cai, Y., Huang., Q.: Quality smoothing for FEC-based Multiple Description Coding, Proc. PCS, San Francisco, CA, USA, Dec. 2004
10. Girod, B., Stuhlmüller, K.W., Link, M., Horn, U.: Packet Loss Resilient Internet Video Streaming. Proc. VCIP, vol. 3653, Proc. SPIE. Jan. 1999. 833–844
11. Kim, B.J., Xiong, Z., Pearlman, W.A.: Low Bit-rate Scalable Video Coding with 3-D Set Partitioning in Hierarchical Trees (3-D SPIHT). IEEE Trans. on Circuits and Systems for Video Technology, 2000, 10(8): 1374–1387

Semi-fragile Watermarking Based on Dither Modulation

Jongweon Kim[1], Youngbae Byun[2], and Jonguk Choi[12]

[1] College of Computer Software and Media Technology, Sangmyung University
7, Hongji-dong, Jongno-gu, Seoul, 110-743, Korea
{jwkim,juchoi}@smu.ac.kr
[2] MarkAny Inc.
10F, Ssanglim Bldg., 151-11, Ssanglim-dong, Jung-gu, Seoul, 100-400, Korea
byunyb@freechal.com, juchoi@markany.com

Abstract. In this paper, a semi-fragile watermarking algorithm is proposed to prevent images from being altered, based on dither modulation and linearity of discrete cosine transform (DCT). As the algorithm transforms DCT coefficients in spatial domain using dither modulation and the linearity of DCT, it embeds watermarking information very fast without DCT. The robustness of the pro-posed algorithm against compression proved in a serious of experiments. Furthermore, because a semi-fragile watermark can be embedded into DCT coefficients with only additions and subtractions, watermarking information can be embedded fast with simple computations. As a result, semi-fragile watermark-ing information can be added in a real time to prevent forgeries and alteration in portable devices with limited resources such as digital cameras, digital camcorders and cellular phones.

1 Introduction

With the wide spread use of internet and a rapid increase of internet users, electronic commerce (EC) enables internet users actively to trade consumer products and tangible goods through networks. As electronic commerce is activated, the certification of digital documents is emerging as an essential procedure for trading and on/offline transactions of products.

In addition, images taken using devices such as digital cameras and digital cam-corders are growing more important. Because these digital contents are digital, they can be easily forged or altered with various types of edition tools, which requires technologies to prevent illegal duplication, protect the copyright and proprietorship of multimedia contents, and to determine if digital contents have been forged or altered.

Technologies have been developed to prove the integrity of digital contents by detecting malicious forgeries and alterations in digital contents using watermarking.

Because the size of an image is quite large, however, various compression methods have been developed and used to store a large volume of image data.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 721–730, 2004.

Because the fragile watermarking technology regards compression as forgery and alteration, it cannot be applied to detect forgeries and alterations in compressed images.

To detect forgeries and alterations in compressed images, we use the semi-fragile watermarking technology. Although it is a fragile watermarking technology, it is robust against compression.

Typically, to be robust against JPEG compression, an image is compressed at a high quantization factor and watermarking information is inserted [1,2]. Because it has to go through discrete cosine transform (DCT) and inverse discrete cosine transform (IDCT) the method requires a lot of computation. Furthermore, to be robust against compression, an image has to be pre-quantized at a high quantization factor and, consequently, it suffers a lot of damage. Another method is to insert watermarking information into a low frequency band because information in high frequency bands is destroyed by image compression [3]. This reduces the volume of computation but the image still suffers a lot of damage like the previous method because watermarking information is inserted to a large number of DCT coefficients.

The present paper reduced computation necessary for watermark insertion by inserting watermark information into some DCT coefficients selected at random using the linearity of DCT, and reduced damage on images by inserting watermark with low strength into target DCT coefficients [6].

2 Watermarking Algorithm Using Dither Modulation

Inserting data through quantization uses two or more codebooks to represent various different symbols. Codebooks can be expressed as scalar quantization of an even or odd quantization index. The most typical structure of watermarking based on quantization is quantized index modulation and quantized projection [4].

This paper uses dither modulation, which is a type of quantized index modulation, to insert semi-fragile watermarking information. The information insertion procedure of the method uses three sets of parameters as in Eq.(1).

$$\begin{array}{l} Set\ 1 : \{\Delta_k\}_{k=1:L} \\ Set\ 2 : \{\underline{d}_1(k)\}_{k=1:L} \\ Set\ 3 : \{\underline{d}_2(k)\}_{k=1:L} \end{array} \tag{1}$$

where, δ_k indicates the quantization level of kth signal, and $\underline{d}_1$ and $\underline{d}_2$ are pseudo noise sequence equivalent to '0' and '1' respectively. In addition,

$$|\underline{d}_1(k)|, |\underline{d}_1(k)| < \frac{\Delta_k}{2}, \quad k = 1,\ 2, \cdots, L \tag{2}$$

The original data signal is divided into segments of the length of L. A bit is inserted to each segment. The modulation method for the kth element of the ith segment is as follows [4].

$$y_i(k) = Q(\underline{x}_i(k) + \underline{d}_i(k)) - \underline{d}_i(k), \quad k = 1,\ 2, \cdots, L \tag{3}$$

where, $Q(\cdot)$ means quantization using quantization level Δ_k. In addition, $\underline{d}_i \in \{\underline{d}_1, \underline{d}_2\}$. As mentioned in [4], detection threshold is decided by the value of $\{\Delta_k\}$.

The procedure of watermark detection is as follows.

Step 1. Add $\underline{d}_1$ and $\underline{d}_2$ respectively to modulated segments
Step 2. Quantize at quantization level $\{\Delta_k\}$ corresponding to each segment
Step 3. Measure the entire quantization error for each case obtained from Step 1
Step 4. Identify the bit with the smallest quantization error as the bit inserted

For the robustness of watermarking, $\underline{d}_2(k)$ is expressed as the operation of $\underline{d}_1(k)$ and Δ_k .

$$\begin{aligned} \underline{d}_2(k) &= \underline{d}_1(k) + \tfrac{\Delta_k}{2}, \quad if\ \underline{d}_1(k) < 0 \\ &= \underline{d}_1(k) - \tfrac{\Delta_k}{2}, \quad if\ \underline{d}_1(k) \geq 0 \end{aligned} \tag{4}$$

3 Semi-fragile Watermarking

3.1 Insertion of Semi-fragile Watermarking

We insert semi-fragile watermark information using dither modulation mentioned in section 2 and the linearity of DCT mentioned in [6]. In particular, as discussed in [6], we can insert watermark information into DCT coefficients of an image just through additions and subtractions in a spatial domain using the linearity of DCT. That is, we can change DCT coefficients selectively by inserting semi-fragile watermark information.

In section 2, we introduced a method of inserting watermarking using dither modulation. However, the method this paper proposes inserts not certain information but semi-fragile watermarking into an image to detect forgeries and alterations, so it uses only quantization different from dither modulation introduced in section 2. That is, in case data are in between A and B as in Figure 1, if pseudo noise sequence is ‘0’ it is quantized to A and if it is ‘1’ it is quantized to B.

DCT coefficient to change the ith block $F_{W_i}(m,n)$ is obtained through Eq. (5).

$$\begin{aligned} &if\ pn_i = 0,\ then\ F_{W_i}(m,n) = Q_\Delta(F_i(m,n) + \tfrac{\Delta}{2}) \\ &if\ pn_i = 1,\ then\ F_{W_i}(m,n) = Q_\Delta(F_i(m,n)) + \tfrac{\Delta}{2} \end{aligned} \tag{5}$$

where, $Q_\Delta(\alpha)$ is:

$$Q_\Delta(\alpha) = \Delta \times \lfloor \frac{\alpha}{\Delta} \rfloor \tag{6}$$

and, is the largest integer that is not bigger than b, namely, Gaussian number.

The procedure to insert semi-fragile watermarking is as follows:

Step 1. Divide an original image into 8 x 8 blocks.
Step 2. Determine the coefficient of discrete cosine domains modified using “pseudo noise sequence 1”
Step 3. Obtain the value of the DCT coefficient to be inserted. To obtain the DCT value of the selected coefficient, DCT must be executed

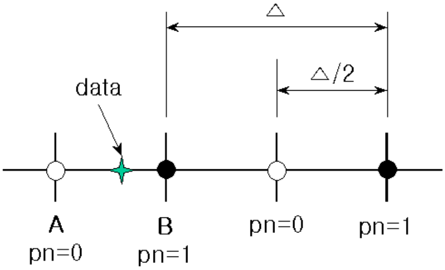


Fig. 1. Dither modulation for semi-fragile watermarking

Step 4. Decide the size of watermarking to be inserted by determining the dither modulation value of the block using "pseudo noise sequence 2"

Step 5. In the spatial domain, multiply spatial domain data corresponding to the DCT coefficient by the volume of insertion obtained in Step 4 and add the results to the original image data

Step 6. Repeat Step 2 through Step 5 for all blocks

Because the size of a block in Step 2 is 8 × 8, the number of pseudo noise bits corresponding to a coefficient is six. The first three bits indicate the abscissa and the next three bits indicate the ordinate. If the pseudo noise sequence corresponding to a block to process is "100110," (5, 7) is selected as the DCT coefficient to be changed. Thus 'pseudo noise 1' is composed of 12 bits per block

To obtain the value corresponding to the DCT coefficient selected in Step 2, DCT is executed in Step 3. Because DCT is independent from each coefficient, it is not necessary to compute the values of all coefficients to obtain the values of selected coefficients. Therefore, DCT is executed only for the selected coefficients. In this study, semi-fragile watermarking information was inserted only to two DCT coefficients, one for the low frequency band and the other for the medium frequency band, for effective detection because values in the high frequency band are removed when loss compression like JPEG compression is executed and in order to reduce damage on the image. That is, DCT was executed only for two coefficients. Because there are a total of 64 coefficients, only 1/32 of them need computation for DCT to insert watermarking.

Step 4 obtains the coefficient of the low frequency band $F_{L_i}(m_L, n_L)$ and that of medium frequency band $F_{M_i}(m_M, n_M)$, and then obtains a and b, the differences between them and $F_{O_i}(m_L, n_L)$ and $F_{O_i}(m_M, n_M)$ in the original image.

$$\begin{aligned} \alpha &= F_{O_l}(m_L, n_L) - F_{L_l}(m_L, n_L) \\ \beta &= F_{O_l}(m_M, n_M) - F_{M_l}(m_M, n_M) \end{aligned} \tag{7}$$

Add the data of spatial domains to the original image using α and β obtained in this way.

$$f_{WM}(x, y) = f_O(x, y) + \alpha f_L(x, y) + \beta f_M(x, y) \tag{8}$$

where,

$f_{WM}(x,y)$: image data to which semi-fragile watermark has been inserted
$f_0(x,y)$: data of the original image
$f_L(x,y)$: data of the spatial domain corresponding to selected coefficient 1
$f_M(x,y)$: data of the spatial domain corresponding to selected coefficient 2
α : variation of coefficient 1
β : variation of coefficient 2

The volume of semi-fragile watermarking inserted is decided according to the size of Δ, the quantization level. Accordingly, Δ is the insertion strength of semi-fragile watermarking.

3.2 Detection of Forgeries and Alterations

Just as in inserting semi-fragile watermark information, so we inspect the forgery and alteration of an image by determining if it was forged or altered using the values of DCT coefficients obtained from 'pseudo noise sequence 1' and 'pseudo noise sequence 2'.

The procedure to detect forgeries and alterations is as follows.

Step 1. Divide the original image into 8 × 8 blocks
Step 2. For each block, identify DCT coefficients to which semi-fragile watermarking information has been inserted using 'pseudo-random sequence 1'
Step 3. For the coefficient of each block obtained in Step 2, compute the probabilities of True and False using 'pseudo-random sequence 2'
Step 4. Decide whether the image has been forged/alternated or not by adding probabilities corresponding to n blocks.

Like in inserting semi-fragile watermarking information, in Step 2 DCT is executed only for selected coefficients.

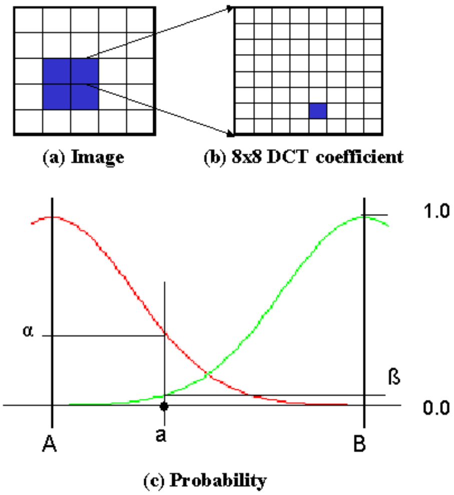
Probabilities computed in Step 3 represent the absolute values of differences as in Eq.(9) and (10). That is, in case the value of 'pseudo-random sequence 2' correspond-ing to each block is '0', the absolute value of the difference between the value obtained from Eq. (5) and the coefficient obtained from Step 2 is the probability of True. In case it is '1', it is the probability of False.

$$\begin{aligned} if\ pn_i = 0,\quad P_{T_i} &= |D_i(m_L,n_L) - F^0_{L_i}(m_L,n_L)| + |D_i(m_M,n_M) - F^0_{M_i}(m_M,n_M)| \\ P_{F_i} &= |D_i(m_L,n_L) - F^1_{L_i}(m_L,n_L)| + |D_i(m_M,n_M) - F^1_{M_i}(m_M,n_M)| \end{aligned} \tag{9}$$

$$\begin{aligned} if\ pn_i = 1,\quad P_{T_i} &= |D_i(m_L,n_L) - F^1_{L_i}(m_L,n_L)| + |D_i(m_M,n_M) - F^1_{M_i}(m_M,n_M)| \\ P_{F_i} &= |D_i(m_L,n_L) - F^0_{L_i}(m_L,n_L)| + |D_i(m_M,n_M) - F^0_{M_i}(m_M,n_M)| \end{aligned} \tag{10}$$

where, L indicates the low frequency band and i indicates the number of a 8×8 block. In addition, $F^0_{L_i}(m_L,n_L)$ indicates the value obtained from dither modulation of a DCT coefficient using Eq. (5) in case pseudo noise sequence is '0', and $F^1_{L_i}(m_L,n_L)$ in case pseudo noise sequence is '1'.

For example, in case the coefficient is a as in (c) of Figure 2, let's say it is quantized into A if pseudo noise sequence 2 is '0' and to B of pseudo noise sequence 2 is '1'. In addition, if the value of 'pseudo noise sequence 2' correspond-ing to the block is '0', a is the probability of True and b is that of False. These


(a) Image

(b) 8x8 DCT coefficient

(c) Probability

Fig. 2. Forgery/alteration detection method

probabilities are used in determining whether the image has been forged/altered or not. That is, forgery/alteration is determined as the following Equation using the probability of True (P_t) and that of False (P_f), which are obtained in Step 3 using Eq. (11) in Step 4 of the forgery/alteration detection procedure.

$$IF \sum_{k=1}^{n} P_{t_k} \geq \sum_{k=1}^{n} P_{f_k}, \quad \begin{array}{l} THEN\ the\ blocks\ is\ no\ forgery \\ Otherwise,\ the\ block\ is\ forgery \end{array} \tag{11}$$

where, n means the number of blocks used to determine forgery/alteration among the blocks divided in Step 1. In this paper, it is 4. That is, forgery/alternation is determined using four 8 × 8 blocks, which neighbor with one another.

4 Experiment Results

In this paper, semi-fragile watermark information was inserted to two DCT coefficients, one for the low frequency band and the other for the medium frequency band.

Table 1 shows PSNR according to the insertion strength of semi-fragile watermarking and performance according to JEPG compression. This experiment used Lena image of 512 × 512 pixels. Quality Factor (QF) in Table 1 represents the degree of JPEG compression. That is, the smaller QF is the stronger the compression is. QF in Table 1 indicates a degree, at which the detection error rate of semi-fragile watermarking information is not higher than 1%. In other words, the detection error rate is less than 1% up to QF in Table 1.

Table 1. Performance according to the insertion strength of semi-fragile watermarking (Δ)

Strength(Δ)	6	8	10	20	30	40	50	60
PSNR[dB]	61.82	57.82	55.64	50.68	46.97	44.48	42.48	40.94
QF	99	82	78	56	31	20	16	14

Figure 3 shows semi-fragile watermark detection errors according to JPEG compression when semi-fragile watermarks have been inserted into 15 images of 1280 × 1024 pixels at insertion strength of 20. Here, the average PSNR was 50.45dB. As mentioned in section 4, better PSNR was resulted because semi-fragile watermark was inserted into only a part of the DCT coefficients.

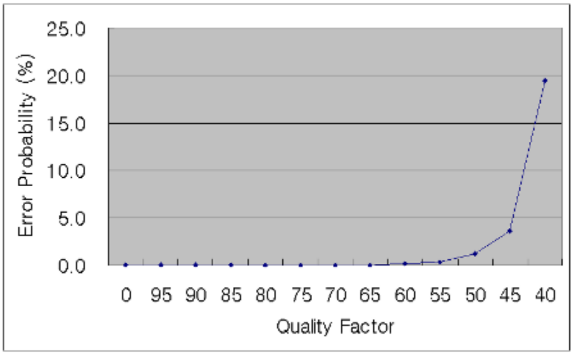


Fig. 3. Robustness against JPEG compression

Although semi-fragile watermark information was inserted at PSNR as weak as 50dB, the result of detection for JPEG compression was satisfactory as Figure 4. When the quality factor (QF) was 55 the average detection error rate was 0.31%. In addition, when QF was 50 it was 1.19%, and when QF was 45 it was 3.63%. However, when QF was 40 the average detection error rate was 19.43%.

When the insertion strength (Δ) was 20, the detection error rate was less than 1% at QF of up to 55. Thus, this study concludes that the semi-fragile watermarking algorithm proposed in this paper can be utilized in proving the integrity of images through detecting semi-fragile watermark information inserted at QF of up to 55.

Table 2 shows semi-fragile watermark detection error rates according to different semi-fragile watermark insertion strengths (Δ) for the image in Figure 4(a). Although PSNR at insertion strength of 60 was 41dB the detection error rate was 0.0% at JPEG compression QF of up to 25, and when QF was 15, 0.05% of semi-fragile watermarking was not detected.

In Figure 4, when image (a), to which semi-fragile watermarking information was inserted, was altered to (b), the altered part was correctly detected as in (c).

Table 2. Semi-fragile watermarking detection error rates for JPEG compression (%)

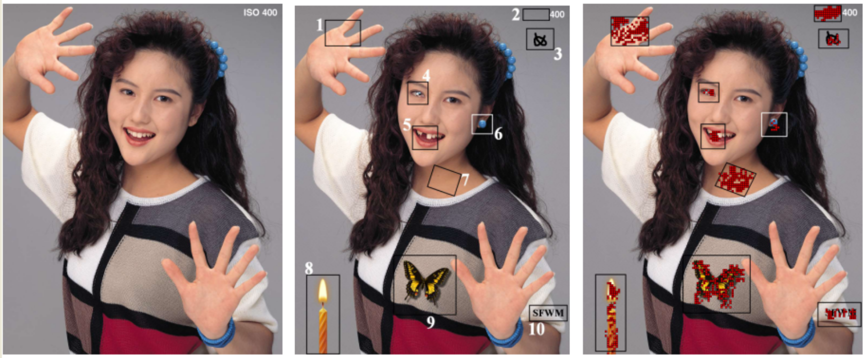
Δ	PSNR	Quality Factor								
	(dB)	95	85	75	65	55	45	35	25	15
10	5.28	0.00	0.07	1.40	19.92	-	-	-	-	-
20	50.45	0.00	0.00	0.00	0.02	0.00	0.41	22.08	-	-
30	46.73	0.00	0.00	0.00	0.00	0.00	0.00	0.03	17.91	-
40	44.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	-
50	42.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14.93
60	40.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05

The size of the original image was 1024 × 1280 pixels, 24 bpp, 3.75 MB, and the size of the altered image, which was compressed through JPEG compression at QF of 60, was 156 KB.

Images were forged and altered in following ways:

1. Paste original image
2. Delete (fill background textures)
3. Add a line drawing
4. Change color (the pupil of the eye)
5. Delete guts (a tooth)
6. Copy (an earring)
7. Rotate
8. Paste another content
9. Paste another contents
10. Replace by computer generated texts

Among the total of 20,480 blocks (one block = 8 × 8 pixels), 1,159 blocks were forged/altered, of which 894 blocks were detected.



(a) Original Image (b) Modified Image (c) Authentication Result

Fig. 4. Forgery/alternation detection results

Because the method proposed in this study inserts semi-fragile watermarking in-formation using the linearity of DCT, it inserts the information very quickly.

In addition, because it inserts only to certain DCT coefficients selected through dither modulation, it maintains high image quality and, at the same time, is robust against compressionpe. That is, as presented in Table 1 and 2, even at PSNR of 50dB the detection error rate was less than 1.0% at QF of up to 55. Considering that semi-fragile water-marking information should be inserted real-time in portable devices with limited resources, methods such as that proposed by Lin[3] may be used. However, Lin's method has a problem in setting the threshold value. That is, the threshold value has to be different according to image and the degree of JPEG compression [3]. Further-more, according to results presented in Lin's paper [3], when the threshold value is set at 0.1 about 75% of blocks are detected at QF of 60. Compare to this, with the method proposed in this paper, when watermarking information was inserted at PSNR of 40 45dB, which was similar to that in Lin's method, the detection error rate was 0.0% at QF of 60 in all of 16 images tested.

5 Conclusions

For semi-fragile watermarking robust against compression, this paper proposed a method of inserting semi-fragile watermark information to certain DCT coefficients selected at random, through additions and subtractions at spatial domains using the linearity of DCT without executing DCT and IDCT. It also used probabilities to correct errors occurring in watermarking detection due to compression and quantization.

The proposed semi-fragile watermark method can perform forgery/alteration detection robust against JPEG compression even at high PSNR. When the insertion strength of semi-fragile watermark (*Delta*) is 20 and PSNR is 50dB, the watermark detection error rate is less than 1% at QF of up to 55. This means that the method can be used to detect forgeries and alterations at QF of up to 55.

The method in this paper can insert semi-fragile watermarks into DCT coefficients quickly and with a small volume of computation just through additions and subtractions at spatial domains. Thus, it can be used to insert semi-fragile watermark information to prevent forgeries and alteration at portable devices with limited resources such as digital cameras, digital camcorders and cellular phones.

Acknowledgements. Solideo Gloria. This work was supported by National Research Laboratory project (2000N-NL-01-C-286) of Ministry of Science and Technology in Korea.

References

1. C.-Y. Lin and S.-F. Chang, "Semi-Fragile Watermarking for Authenticating JPEG Visual Content," Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, Cali-fornia, pp. 140–151, January 2000

2. Kurato Maeno, et al, "New Semi-Fragile Image Authentication Watermarking Techniques Using Random Bias and Non-Uniform Quantization", Proc. SPIE Security and Watermark-ing of Multimedia Contents, San Jose, California, pp. 659–670, January 2002
3. Eugene T. Lin, et al. "Detection of image alterations using semi-fragile watermarks", Proc. SPIE Security and Watermarking of Multimedia Contents, San Jose, California, pp. 23–28, January 2000
4. Chen B., and Wornell G.., "Quantization Index Modulation, a class of provably good meth-ods for digital watermarking and information embedding", IEEE Trans. On Info. Th., Vol. 47, No. 4, pp. 1423–1443, May 2001
5. Chen B., and Wornell G.., "Dither modulation: a new approach to digital watermarking and information embedding", Proc. SPIE conference on security and watermarking multimedia contents, pp. 342–353, January 1999
6. Jongweon Kim, Youngbae Byun, Jonguk Choi, "Image Forensics Technology for Digital Camera", PCM2004 accepted, Nov., 2004
7. Eugene T. Lin, et al. "Detection of image alterations using semi-fragile watermarks", Proc. SPIE Security and Watermarking of Multimedia Contents, San Jose, California, pp. 23–28, January 2000

An Adaptive Steganography for Index-Based Images Using Codeword Grouping

Chin-Chen Chang[1], Piyu Tsai[2], and Min-Hui Lin[3]

[1] National Chung-Cheng University, Chiayi Taiwan 621,ROC
[2] National United University, Miaoli Taiwan 360, ROC
[3] Providence University, Taichung Taiwan 433, ROC

Abstract. Hiding secret messages into index-based images is difficult because it suffers from an amount of image degradation and a limited hiding capacity. In this paper, an adaptive steganographic scheme for index-based images is proposed. The codewords are grouped into sub-clusters according to the relationship among codewords. Also, the size of the sub-cluster determines the hiding capacity of the proposed scheme. The experimental results show the performance of the proposed steganography. In comparison with the least significant bit modification method (LSB), a better stego-image quality is obtained by the proposed scheme. In comparison with Jo and Kim's and Fridrich's methods, a higher hiding capacity is provided by the proposed scheme while the near image quality remains. Furthermore, an adaptive hiding capacity is achieved by the proposed method.

1 Introduction

With the digitalization of data and the networking of communication, communication security over the Internet is becoming more and more crucial [1]. Basically, the Internet is an open channel and security problems such as modification, interception, as well as others, usually exist. Several different approaches have been proposed to make private communication secure [2]. The first approach encrypts the secret message to prevent information from leaking out. In such schemes, the secret message is protected by transforming it into an unrecognizable form. Only the authorized user can retransform it back to its original form by using secret information. Many famous encryption schemes, such as RSA, DES, and so on are widely used in the commercial market. However, the meaningless form could leave a clue and inspire an unauthorized user to explore the original message.

Another approach, called steganography, hides a secret message in a widespread cover material to avoid suspicion. The concept of steganography is similar to the concept of camouflage, which is used by some animals to protect themselves from being attacked. Several steganographic schemes have been developed to solve the privacy problem [3,4,5,6,7,8]. In Lee and Chen's method [3], the least significant bit (LSB) of each pixel in the cover image is modified to embed the secret message. In Tsai et al.'s scheme [4], the bit plane of each block truncation coding (BTC) block is exploited to embed a secret message. In

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 731–738, 2004.

Chang et al.'s method [5], the middle frequency coefficients of the DCT transformed cover image are employed to embed the secret message. In Spaulding et al.'s method [6], the embedded zerotree wavelet (EZW) encoded cover image is used to embed the secret message. The bit-plane complexity segmentation (BPCS) and the visual system are explored to determine the hiding capacity and the stego-image quality.

Only a few methods work on the index-based cover images [7,8]. In fact, the index-based images such as vector quantization (VQ), color quantization (CQ)-based (palette-based) images, are widely applied. Currently, palette-based images such as GIF files have been widely used on the Internet and web pages such that all browsers can recognize them. In 2002, Jo and Kim proposed a watermarking method based on VQ [7]. In their scheme, the codewords in the codebook are partitioned into three sub-clusters. The higher similarity of members between two special sub-clusters is preserved. This feature is employed to hide the watermark information. In Fridrich's method [8], the parity of the searched color in the palette is examined to match the embedding message. In these index-based methods, the problems of the image quality degradation of the stego-image and the hiding capacity limitation of the cover image occur.

To remain at an acceptable stego-image quality, the hiding capacity is usually sacrificed, and vice versa. To conquer the problems described above, in this paper, we shall propose an index-based steganographic scheme in which the hiding capacity is adaptive and the good stego-image quality remains. To achieve our goal, the relationship among codewords is exploited. The rest of this paper is organized as follows. In Section 2, Jo and Kim's technique is briefly described. Next, the proposed steganography is introduced in Section 3. In Section 4, the experimental results of the proposed scheme are shown. Finally, some conclusions are given in Section 5.

2 An Overview of Related Work

In index-based image hiding, the least significant bit (LSB) modification for each searched index is simple and applicable. In Jo and Kim's watermarking, all codewords are grouped to hide the watermark information. To clear the proposed method, Jo and Kim's technique is briefly described.

In Jo and Kim's technique [7], all codewords of the codebook are grouped into three groups (G_{-1}, G_0, G_1) according to the similarity between codewords. The group G_{-1} consists of codewords, which are unsuitable for embedding the watermark information. The other two groups consist of codewords in which each codeword in one group corresponds to a similar codeword in the other group. In other words, each codeword in the group G_0 must have another relative codeword in the group G_1, and their similarity is high. Both groups G_0 and G_1 can be considered, respectively, to represent the bit values of 0 and 1 in the watermark embedding.

For each block embedding, the closest codeword is first searched and then, the group to which the codeword belongs is determined. If the closest codeword belongs to the group G_{-1}, this codeword remains and no watermark information is embedded. If the group representation matches the watermark information, the closest codeword is also preserved. Otherwise, another similar codeword located in the corresponding group is employed. For example, if the embedding watermark is 0 and the group of the closest searched codeword is G_1, another higher similarity codeword located in the group G_0 is selected. From that, the closest searched codeword may be modified to carry the watermark according to the group representation. However, the hiding capacity in this method is small since each codeword can only carry a bit of the watermark information at most.

3 The Proposed Scheme

In this section, the proposed steganography based on the codeword grouping will be introduced. First, the codeword grouping is described. Then, the proposed embedding and extracting procedures are introduced. Finally, the quality, capacity and security are discussed.

3.1 The Codeword Grouping

Generally, a set of codewords, called codebook/palette, is employed in the index-based images encoding/decoding. The codewords are usually generated using the codebook/palette generation algorithm. In the index-based image encoding, the image is first partitioned into blocks of $n \times n$ pixels. Each block encoding finds the closest codeword from the codebook in terms of a similarity measure (ex. squared Euclidean distance, MSE) to represent the encoding block. The index of the closest searched codeword in the codebook is used to represent the encoding block by reducing the storage space. However, each codeword generally represents a set of training blocks. Therefore, any modification of the encoded index may incur a greater amount of image distortion. To reduce the image distortion caused by the index modification, the relationship between codewords is exploited. The square Euclidean distance between codewords is used to indicate the relationship.

The proposed codeword grouping partitions the codewords into different sub-clusters according to the relationship between codewords. The codewords with stronger relationship are grouped into a sub-cluster in which the distance of members is less than a predefined threshold. Particularly, if the relationship between a codeword and others is greater than a threshold, this special codeword may be grouped individually. In other words, this sub-cluster only contains a member of itself.

The number of sub-cluster members determines the hiding capacity in the proposed scheme. The more sub-cluster members there are, the higher the hiding capacity will be. To enhance the hiding capacity, the sub-clusters with larger members will be grouped first. The sub-clusters with smaller members will then be grouped. Finally, the residual codewords with only one-member are generated individually as a sub-cluster. For each sub-cluster, the number of members is restricted to the power of 2. From that, the codewords are grouped into many sub-clusters with different members. The result of the grouping is determined by the relationship of the codewords and the grouping thresholds.

For example, a codebook with 256 codewords will be grouped into 4-member, 2-member and 1-member sub-clusters according to grouping thresholds. First, 4-member sub-clusters are selected with the distance is less than $TH4$. Then, 2-member sub-clusters are selected with the distance is less than $TH2$. In the sub-cluster grouping, a codeword may preserve a higher relationship with many codewords. To keep a codeword only belonging to one sub-cluster, the distance with the least is first grouped as a sub-cluster. Next, the codewords holding the second least distance form another sub-cluster, which are members of this sub-cluster, are never taken into other sub-clusters. From that, the most similar codewords can be grouped as a sub-cluster. Also, the similarity of members in a sub-cluster is always restricted to a predefined threshold.

Finally, each residual codeword is grouped as a single member sub-cluster individually. After that, the codewords are grouped into different member sub-clusters.

3.2 The Embedding Procedure

In the proposed embedding procedure, the sub-cluster to which the closest searched codeword belongs is first identified. And then, the original encoded codeword is modified to hide the secret message. Therefore, it is easily incorporated into the original encoding procedure.

Before the cover image encoding is performed, the codewords have been grouped into different member sub-clusters. In the encoding procedure, each encoding block finds the closest codeword to represent this block. Once the closest codeword is found, the sub-cluster to which the closest searched codeword belongs is detected. The number of this sub-cluster's members is calculated. If the number of sub-cluster members is greater than one, the embedding procedure is triggered. Otherwise, this block is unsuited to hide the secret message.

In the embedding procedure, the number of sub-cluster members indicates how many bits of the secret message can be embedded. In other words, the number of sub-cluster members determines the hiding capacity of the embedding block. For example, if the number of sub-cluster members is n, log_2 n-bit secret message can be embedded into. To embed the log_2 n-bit secret message, the index of the closest searched codeword may be modified. This modification is determined by the secret message to be embedded and the order of members in the corresponding sub-cluster. The member whose order matches the embedded secret message is adopted to replace the closest searched codeword. In other words, the index of the new codeword is used to encode the encoding block. From that, the encoding block is modified and the secret message is embedded.

An example shown in Table 1 illustrates the proposed embedding procedure. In Table 1, a sub-cluster consists of four members, which are ordered as 0, 1, 2, and 3, respectively. The indices of these four members in the codewords are 100, 130, 107 and 90, respectively. In the block encoding, the index of the closest codeword is 130 and the number of the identified sub-cluster members is 4. The

4-member sub-cluster indicates that a 2-bit secret message can be hidden. If the 2-bit secret message to be embedded is valued at 3 (11 in binary), an index value of 90, which is ordered 3 in the sub-cluster, is used to encode this block. In other words, the closest searched index 130 is replaced by an index of 90.

3.3 The Extracting Procedure

In the extracting procedure, the sub-cluster to which each decoding index belongs is first identified. If the number of the identified sub-cluster members is greater than one, the current decoding index hides a portion of the secret message. Otherwise, no more secret message is embedded. To extract the hidden message, the number of the sub-cluster members, which indicates the size of the hidden message, is calculated. Then, the order of the decoding index in the sub-cluster is the secret message. For instance, if the decoding index is 90, its relative members of the corresponding sub-cluster are shown in Table 1. The number of the sub-cluster members and the order of the decoding index are 4 and 3, respectively. A 2-bit secret message valued at 3 (11 in binary) is extracted.

Table 1. The index and the order of sub-cluster members

Sub-cluster order	0	1	2	3
Codeword index	100	130	110	90

3.4 Quality, Capacity, and Security Considerations

In the image-hiding schemes, there is a tradeoff between the quality of the stego-image and the capacity of the cover image. To preserve a higher stego-image quality, the hiding capacity is usually sacrificed, and vice versa. In the proposed scheme, the codeword grouping provides a possible solution to solve it.

In the codeword grouping, a set of thresholds is used to determine the members of a sub-cluster. Generally, the larger the threshold is, the more sub-clusters with large size there will be. The large size of some sub-clusters provides a high hiding capacity and causes a large degradation of the stego-image. Therefore, choosing an adequate threshold is quite important. Once choosing an adequate threshold, a certain part of a secret message can be embedded, and the image quality of the stego-image would be acceptable.

On the security consideration, the secret message can be permuted before it is embedded. In addition, the sequence of the embedding pixels can also be reorganized using a pseudo random number generator in which the secret key is hold and security is enhanced.

4 Experimental Results

Color quantization (CQ) encoder was simulated to evaluate the performance of the proposed scheme. Three RGB color images, “Lena,” “Sailboat,” and “Logo” of 512 × 512 pixels were encoded by the color quantization technique. The palette of each color image with 256 3-dimensions was generated by Photoshop version 6.0 with optimal option. The images shown in Figure 1 were taken as cover images. Three binary images, “CCU,” “IEEE,” and “NUU” of 128 × 128 bits shown in Figure 2 were used as secret messages.

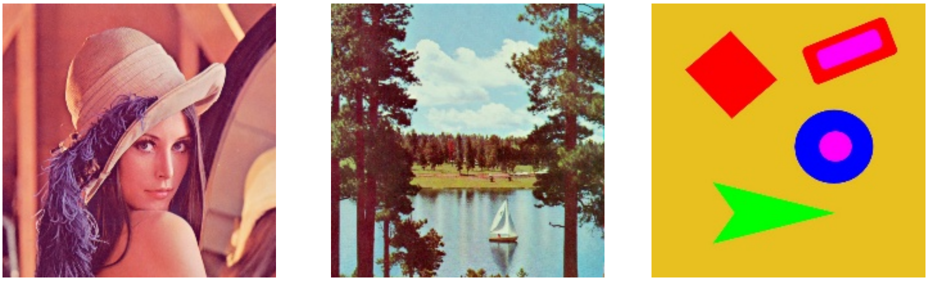


Fig. 1. The palette cover images of “Lena”,“Sailboat” and “Logo”

Fig. 2. The secret images of “CCU”, “NUU” and “IEEE”

To determine the hiding capacity, the relationship between codewords and the characteristics of images are explored. Table 2 shows the hiding capacity of cover images according to the different grouping threshold. The capacity was computed according to the number of blocks, which belonged to 4-member or 2-member sub-clusters. From the hiding capacity shown in Table 2, it is clearly seen that the capacities were different. The capacity in the second column was higher than that in the first column because the larger grouping threshold was employed in the second column. For images, “Lena,” “Sailboat,” and “Logo”, different hiding capacities were provided in the same grouping threshold. This is because the characteristics of the cover images were different. From Table 2, it is noted that the adaptive hiding capacity could be achieved by using different grouping threshold.

To evaluate the performance of the proposed steganographic scheme, the quality of the stego-image by the proposed method was measured. The experimental results are shown in Table 3 in which the cover images and the secret images of 128 × 128 bits were simulated and the first column grouping results

Table 2. The numbers of different member sub-cluster and total capacities for color images with different thresholds

	Sub-cluster numbers			Sub-cluster numbers		
	4-member ($TH4$=500)	2-member ($TH2$=300)	Capacity	4-member ($TH4$=1000)	2-member ($TH2$=500)	Capacity
Lena	145460	64668	355588	185800	42129	413729
Sailboat	0	169465	169465	139129	61077	339335
Logo	30220	208124	268564	30890	207939	269719

Table 3. The quality of stego-images by CQ encoder and the proposed method PSNR Images CQ Proposed method

	CQ	Proposed method		
		CCU	IEEE	NUU
Lena	36.7206	36.3387	36.3298	36.3390
Sailboat	33.0924	32.8186	32.8475	32.8445
Logo	72.8156	67.4746	67.6857	66.7135

Table 4. The image quality and hiding capacity by LSB, Jo and Kim's, Fridrich's, and the proposed method

	LSB		Jo and Kim's		Fridrich's		Proposed method	
	PSNR	Capacity	PSNR	Capacity	PSNR	Capacity	PSNR	Capacity
Lena	30.5022	262144	36.4548	206248	36.4513	262144	36.3387	355588
Sailboat	29.8380	262144	32.8186	167696	32.5419	262144	32.8186	169465
Logo	37.0856	262144	68.5510	237896	68.3362	262144	67.4746	268564

shown in Table 2 were used. From Table 3, it can be seen that the stego-image quality was near the CQ encoded image quality. In other words, the stego-image distortion was less and unnoticed. It is noted that the difference of PSNR values in the image "Logo" between CQ and the proposed method was large, but the stego-image quality remained a higher PSNR value. So, it is almost undistinguished by the human eyes. From this Table, it is shown that the proposed method provided a good stego-image quality.

A comparison of the stego-image quality and the hiding capacity among LSB, Jo and Kim's, Fridrich's and the proposed method were also performed. The secret images "CCU" 128 × 128 bits were embedded. In Table 4, a better image quality and a higher hiding capacity were achieved in comparison with LSB in images "Lena" and "Logo". Also, the near image quality and a higher hiding capacity were obtained in comparison with Jo and Kim's, and Fridrich's. In the cover image "Sailboat", the proposed method preserved the image quality better than LSB and Fridrich's method while providing the least hiding capacity. From this Table it demonstrates that the proposed method provides adaptive hiding capacity and an acceptable image quality remains.

5 Conclusions

In this paper, a codeword grouping-based steganographic scheme for index encoding images has been presented. The relationship of codewords is explored to group different member sub-clusters. The size of the sub-cluster determines the hiding capacity. The experimental results have shown that adaptive steganography is achieved. Meanwhile, the degradation of the stego-image is less and nearly unnoticed.

The comparison results also indicate that the proposed method provides a good stego-image quality and supports adaptive hiding capacity better than others. In comparison with Jo and Kim's method, the proposed method provides approximately one and half hiding capacity while preserving near image quality. Both the image quality and hiding capacity are close between Fridrich's and the proposed method. However, there is not an adaptive mechanism in Fridrich's method. To sum up, the proposed scheme can meet different requirements: high image quality of the stego-image or great hiding capacity by adjusting the grouping thresholds.

References

1. Cheng, Q. and Huang, T. S.: An Adaptive Approach to Transform-Domain Information Hiding and Optimum Detection Structure. IEEE Transactions on Multimedia, **3**, (2001) 273–284
2. Artz, D.: Digital Steganography: Hiding Data within Data. IEEE Internet Computing, **5**, (2001) 75–80
3. Lee, Y. K. and Chen, L. H.: High Capacity Image Steganographic Model. Proceedings of IEE International Conference on Vision, Image and Signal Processing, **147**, (2000) 288–294
4. Tsai, P. Hu, Y. C. and Chang, C. C.: An Image Hiding Technique Using Block Truncation Coding. Proceedings of Pacific Rim Workshop on Digital Steganography, Kitakyushu, Japan, (2002) 54–64
5. Chang, C. C. Chen, T. S. and Chung, L. Z.: A Steganographic Method Based upon JPEG and Quantization Table Modification. Information Sciences, **141**, (2002) 123–138
6. Spaulding, J. Noda, H. Shirazi, M. N. and Kawaguchi, E.: BPCS Steganography Using EZW Lossy Compressed Images. Pattern Recognition Letters, **23**, (2002) 1579–1587
7. Jo, M. and Kim, H. D.: A Digital Image Watermarking Scheme Based on Vector Quantization. IEICE Transactions on Information and Systems, **E85-D**, (2002) 1054–1056
8. Fridrich, J.: A New Steganographic Method for Palette-Based Images. Proceedings of the IS&T PICS Conference, Savannah, Georgia, (1999) 285–289

DH-LZW: Lossless Data Hiding Method in LZW Compression

Hiuk Jae Shim and Byeungwoo Jeon

Sungkyunkwan University, 300 Chunchun-Dong Jangan-Gu Suwon, Korea,
waitnual@ece.skku.ac.kr, bjeon@yurim.skku.ac.kr

Abstract. In compressed data, there is not enough room available for data hiding, consequently many data hiding methods try to modify original data before compression. One of well-known lossless compression methods is LZW. It has been widespread due to wide use of GIF format. In spite of its reputation on compression, however, few data hiding methods are directly applicable to LZW itself. This may be due to few redundancies remained in losslessly compressed data. In this paper, we propose the DH-LZW method that embeds a certain message to source data in a lossless manner. Forced update method and an efficient data embedding scheme are proposed to be used. Finally, we show promising experimental results.

1 Introduction

A lot of data hiding methods have been developed as a mean of secret data communication. Accordingly, numerous techniques have been proposed in the name of either steganography or watermarking, which all belong to data hiding techniques in wide sense. In this paper, we are especially interested in developing a lossless data hiding method that can be generally applied to many common lossless compression applications.

Among several approaches in data or image compression, LZW is a well-known technique. Since it refers to a dictionary storing single and their combined symbols, LZW is classified as a dictionary-based technique. It is different from other LZ-family such as LZ77 in defining and handling the matching window. However, they still share common characteristics in that while reading a new symbol (and decompressing codes by decoders), the encoder and its decoder both construct a code table (i.e., dictionary), therefore, there is no need to convey the updated code table to decoder. In addition, the LZ-family is a kind of universal algorithm - data is compressed without any prior knowledge such as the probability density function of source. LZW has been widely used in many applications, for instance, in GIF and TIFF, and PDF writer (in compressing image). However, proposed are few methods which embed user message directly into LZW compressed data. Rather, a number of data hiding methods have been applied to GIF than LZW itself. The reason is that there rarely is redundancy left in LZW compressed data. As a consequence, very few data hiding methods have been built directly on LZW.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 739–746, 2004.

There is two kind of information which can undergo modification in order to hide data in GIF: one is the color definition of palette, and the other is the color index of palette. The parity of palette index can be modified to embed data [1], [2]. Niimi et al. has introduced one way of applying BPCS method to palette-based images. Since BPCS alters one channel of color index, the modified value may not be found in 256 colored palette. In this case, newly generated color definition replaces that of the color index which has the smallest distance from the new one [3]. Ogihara et al. proposed a method which is quite different from others in that they utilized both the value and length of color index [4]. While previous data hiding methods modify palette indexes, value of palette, or color components of source, the compression algorithm itself of LZW has no reason to be left out in modification for data hiding. One way is to match the parity of a prefix code index with that of hiding data: if they match, then the code is unchanged; otherwise, the value of the last symbol in codeword is changed in order to find another parity-matching code in the given dictionary. If the matched code is not found in the dictionary, the code is reduced by one symbol to find a match. One disadvantage of this algorithm is the distortion to the original data caused by data embedding procedure. After data hiding, no severe data distortion may occur in case of 256 color indexes; however it can pose a visually serious problem especially when the original image is represented by only small number of colors such as 32, 16, etc. In this paper, we propose DH (data hiding)-LZW algorithm which causes no distortion to source and is applicable not only to GIF algorithm, but to more wide scope of applications employing the LZW algorithm. Therefore the proposed DH-LZW can be applied to any kind of data only if it is compressed by LZW. The proposed method is described in Section 2, and in Section 3, its experimental results will be shown. We summarize the whole procedure and draw conclusion in Section 4.

2 Data Hiding in LZW Compression

2.1 Overview of Proposed Forced Update

Not like statistical strategies such as Huffman, or the arithmetic coder, etc, the compression process of LZW in GIF is relatively simple. Without any need to investigate the whole probability density of source, LZW is performed as it encounters a new symbol. Moreover, the code table, or the dictionary is constructed dynamically at the decoder side as well as the encoder side, which means that the same procedure of rebuilding their own dictionary is performed at both encoder and decoder. Since dynamical dictionary updating precludes sending additional information describing changes to the dictionary, it provides us new possibility of considering a dictionary as a target domain where some data can be embedded.

Since the main target of compression methods is to find and to reduce the redundancy of given data and then finds the best way to reduce the redundancy, it is not easy to modify the compressing procedure. However, as described in Section 1, there are two different elements in GIF encoder which may undergo modification to embed messages. One is the symbol definition such as a palette, and the other is the length of the symbol, where length means how many single letters are joined together to form the symbol. Modifying the first one in data hiding surely causes distortion to source data, however, changing the second information does not bring distortion if one can actually find a possible way to control the second element properly. Changing symbol definition does not modifies the LZW compressing procedure, therefore it can be thought preprocessing to LZW compression, however changing the length of the symbol directly affects LZW compressing procedure. The change of the symbol length is reducing the length by defining a new symbol by detaching the last letter of the original symbol. When the length is reduced by 1, there always exists a symbol which equals the "reduced symbol" since LZW compression utilizes a previously updated symbol as the prefix of a new symbol about to be updated. Note that reducing the length, or, in a more general term, forced update of an already existing symbol does not bring any compliance problem to the existing LZW algorithm, although it may cause slight decrease in compression performance. We utilize this forced update as means for data hiding. For better understanding of the forced update, the difference between conventional LZW and DH-LZW is illustrated in Figure 1.

LZW			Forced Update in DH-LZW			
input	output	new symbol	input	output	new symbol	Dictionary update
a	a	256=aa	a	a	256=aa	
b	a	257=ab	b	a	257=ab	
b	b	258=bb	b	b	258=bb	
a	b	259=ba	a	b	259=ba	
bb	257	260=abb	bb	257	260=abb	
aa	259	261=baa	aa	259	261=baa	
bba	260	262=abba	bba	260	262=abba	
ba	257	263=aba	ba	257	263=aba	
bbab	262	264=abbab	bbab	260	**264=abba**	**Forced update**
ba	258	265=bba	bab	262	265=abbab	

Fig. 1. The procedures of usual LZW and DH-LZW, input data to compress: abbabbaabbababbabbab

The proposed DH-LZW process (as shown in Figure 1) defines a THD (=4) to skip data hiding for the symbol whose length is less than or equals to THD. However, when a symbol whose length is larger than 4 occurs, data embedding becomes possible. Consequently data hiding procedure is started in accordance with message bits. As shown in Figure 1, symbols having code indexes from 256 to 263 undergo usual updating process since the symbol length is not larger than THD, and symbol with code indexes 264 and 265, where the length meets the condition of THD, undergoes forced update in order to embed message bits. In case of Figure 1, forced update is performed only on the index 264.

Before we outline the decoding process, it is important to understand why the forced update does not affect conventional LZW decompression process. In

LZW				Forced Update in DH-LZW			
input	old code	output	new table entry	input	old code	output	new table entry
a		**a**		a		**a**	
a	a	a	256=aa	a	a	a	256=aa
b	a	b	257=ab	b	a	b	257=ab
b	b	b	258=bb	b	b	b	258=bb
257	b	ab	259=ba	257	b	ab	259=ba
259	257	ba	260=abb	259	257	ba	260=abb
260	259	abb	261=baa	260	259	abb	261=baa
257	260	ab	262=abba	257	260	ab	262=abba
262	257	abba	263=aba	260	257	abb	263=aba
258	262	bb	264=abbab	262	260	abba	264=abba
257	258	ab	265=bba	b	262	b	265=abbab
Decoded : aabbabbaabbababbabbab				Decoded : aabbabbaabbababbabbab			

Fig. 2. Procedures of LZW and DH-LZW decoding

fact, the reason is simple: although the forced update causes redundant symbols in dictionary, LZW compression process uses the earliest updated index as the prefix of the new symbol during searching the match, and in LZW decompression, decoder reads a symbol and updates it in exactly the same manner with compression step. That is, the LZW decoder imitates the encoder's process of updating dictionary, which makes LZW decompression feasible, regardless of the redundant symbol. The difference between a usual LZW decoder and the DH-LZW decoder is illustrated by an example in Figure 2. The columns in the Figure denoted by "input" are from the results of Figure 1. From Figure 2, what we can observe is that the right side DH-LZW decoder de-compresses the input data stream in the exactly same manner as a usual LZW decoder.

2.2 Data Embedding Procedure

As described in previous section, data embedding is made possible by the forced update. The main principle of the forced update is to generate redundant codes in LZW dictionary, consequently the compression performance is decreased. For instance, dictionary updating procedure becomes slower and LZW coder generates an output with one symbol by forced update. Moreover, when we simply utilize the parity of a symbol length as a bit of message, there may be more to lose than gain as the amount of embedding becomes larger. Therefore it is important to find proper data embedding scheme which enables one to hide more than 1 bit, when forced update is performed. We propose one to embed information efficiently.

Since forced update adds redundant codes to dictionary, decoder surely knows the forced updated index. Therefore when we divide the whole size of dictionary into several sub-regions, we can utilize the individual sub-region as an independent message carrier. The length of sub-region is denoted as *r_length* and defined as an integer which is power of 2. This scheme is illustrated in Figure 3.

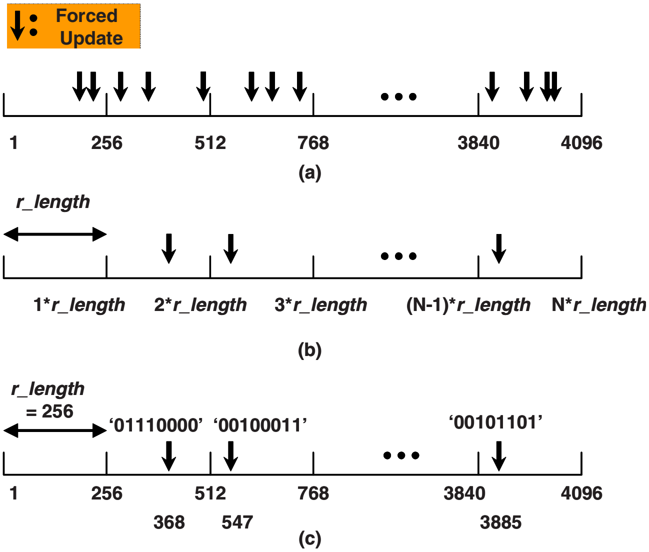


Fig. 3. Difference of data embedding scheme, (a) embedding by parity matching of Forced updated indexes; (b) embedding by sub-region carrier with arbitrary *r_length*; (c) embedding by sub-region carrier with *r_length* = 256

In Figure 3 (a), the numbers from 1 to 4096 are the indexes of LZW dictionary; the largest index may be larger or smaller than 4096 according to applications. Each arrow represents index with forced update. Assuming that message bits are uniformly distributed with '0' and '1' and half of indexes are to be modified according to message bits, total embedded bits are approximately two times larger than the number of forced update. Therefore, when the number of forced update is assumed to be 64, approximately 128 bits are to be embedded.

However, when we divide the whole dictionary indexes into sub-regions as illustrated in Figure 3 (b), we can select and perform forced update on one index within each sub-region. Embedding procedure is as follows. Firstly *r_length* is determined and then message bit string with length *r_length* is read. The index (mod *r_length*) which is the same with message bit string is selected for forced update. Then we can embed message as many as *r_length* bits in every sub-region. Figure 3 (c) shows an example with *r_length* = 256. Since *r_length* is 256, 8bits can be assigned to a sub-region. When message bit string read is '01110000', index (mod *r_length*) can be calculated. For the second sub-region, 368 (mod 256) is 112 and its binary representation is '01110000', therefore index 368 is selected and forced updated. In every sub-region, selected index undergoes forced update in the same way. In this case, total amount of embedding easily calculated by

$$\left(\frac{total\ number\ of\ index}{r\_length} \times log_2\left(r\_length\right)\right) \tag{1}$$

In case of the above example the maximal embedding bits are (4096/256) × 8 = 128. Comparing with Figure 3 (a), the number of forced update is reduced to one quarter; however, total amount of embedding is the same. The reduced number of forced update indicates that the loss of compression efficiency is lowered. In real embedding procedure, the length of the selected index may not be larger than THD. In this case, we have to wait for the next sub-region and check the length again. Therefore, real amount of embedding is less than ideally calculated value. The whole encoding and decoding of DH-LZW process are as follows.

```
Encoding:

 Step.1: Determine a THD and value of r_length which specify the
         shortest symbol length to embed a message and the length
         of sub-region, respectively.
 Step.2: During LZW procedure, read log2(r_length) bits from
         embedding message when index updating procedure is
         entering a new sub-region.
 Step.3: Perform LZW encoding until an index (mod r_length) is
         equal to the read bits.
         If the index is found, go to step 4.
 Step.4: If the index whose length is larger than THD is
         encountered, perform a forced update and go to
         step 2. If the length is not larger than THD,
         noting is done and forced update is executed
         to the next sub-region and go to step 3.
 Step.5: The procedure is repeated from 2 to 4
         until exhausting the message bits.

Decoding:

 Step.1: Process LZW decompression of the compressed data stream.
 Step.2: During the process, if we encounter an index whose
         length is larger than THD, check whether it is
         generated by forced update. If yes, index (mod r_length)
         is performed in order to extract message bits, otherwise,
         repeat the procedure until all the compressed data bits
         are exhausted.
```

3 Experimental Results

For the simulation of image, four well-known images of lena, peppers, splash, and baboon images are considered. They have the same file size in order to observe

Table 1. Maximum amount embedding for each sample image with *r_length* value 128 and 256 with THD = 2 (* : byte is the unit)

File Name	*r_length*	Embedded Bits	Original Size*	LZW*	DH-LZW*	Increased Size(%)
Lena	128	329	65,536	53,889	53,899	0.0186
Splash	128	742	65,536	35,221	35,310	0.2527
Peppers	128	616	65,536	45,150	45,188	0.0841
Baboon	128	203	65,536	58,649	58,682	0.0562
Lena	256	208	65,536	53,889	53,899	0.0186
Splash	256	432	65,536	35,221	35,331	0.3123
Peppers	256	312	65,536	45,150	45,142	-0.0177
Baboon	256	120	65,536	58,649	58,665	0.0273

the different effects about each unique feature of their own. Each size of image file is shown in Table 1.

The proposed DH-LZW algorithm embeds data at the cost of compression efficiency. We can observe several features from Table 1. One is that the size of sample images is very slightly increased with the proposed method; accordingly it shows that the proposed method is efficient for data hiding. Another is that more data can be embedded as the value of *r_length* becomes smaller. This can be apparently shown by equation 1. On the other hand, since smaller *r_length* means increased occurrence of forced update, too small *r_length* value would cause large loss of compression efficiency. Therefore there arise two different figures of merit. One is the maximum amount of payload and the other one is the insensitivity in file size increment after embedding. The maximum payload size and the degree of file size increment can be controlled by setting an appropriate threshold value (THD) and a sub-region length, *r_length* value; however, both can not be achieved at the same time. Therefore we need to determine in advance which direction leads us to more meaningful result. While a mathematical model about payload (or capacity) is required to estimate the possible amount of embedded data, it is not that simple. Since the forced update process alters future patterns to be unpredictable ones, there is only a way of approximating the estimation so far.

4 Discussions

We have introduced a data hiding method in LZW compressed data. The basic concept of the proposed method, DH-LZW, is relatively simple compared to other conventional techniques. In DH-LZW, THD and *r_length* value are defined first, and then the symbol length is utilized during the LZW compression process. The merits of DH-LZW are such that it is a lossless data hiding method, and is compatible with the con-ventional LZW decoders. Therefore massage embedded DH-LZW data can be de-compressed by the general LZW decoders without any conformance problems, and when we need to extract the hidden massage, the embedded message can be retrieved only by DH-LZW decoder. Moreover, it is applicable not only to image data, but to any type of application employing the LZW algorithm, including text, audio, etc.

The proposed DH-LZW algorithm is a data hiding technique that embeds message data into source in a lossless manner. Since LZW compresses source with dynamically constructed dictionary, it is reasonable to handle the dictionary instead of source itself. Handling the dictionary provides a lossless data hiding method, however, another interpretation of embedding data in the dictionary is that it imposes additional redundancy on the dictionary of LZW. That is, as we embed messages, the efficiency of LZW compression would be decreased. Therefore the appropriate analysis of the relationship between the amount of hiding data and compression efficiency is required and it will be our next research topic.

References

1. Fridrich, J., and Du Dui: A new steganographic method for palette-based images IS&T PICS Conference, (1999) 285–389
2. Fridrich, J., and Du Dui: Secure Steganographic Method for Palette Images 3rd Int. Workshop in Information Hiding, (1999) 47–60
3. Niimi, M., Esaon, R.O., Noda, H., and Kawaguchi, E.: A BPCS Based Steganographic Method for Palette-Based Images Using Luminance Quasi-Preserving Color Quantization Proceedings of Pacific Rim Workshop on Digital Steganography, (2002) 84–92
4. Ogihara, T., and Kaneda, Y.,: Data Embedding into Compressed Data of GIF Images Proceedings of Pacific Rim Workshop on Digital Steganography, (2003)

Blind Image Data Hiding in the Wavelet Domain

Mohsen Ashourian[1] and Yo-Sung Ho[2]

[1] Islamic Azad University of Iran, Majlesi Branch,
P.O.Box 86315-111, Isfahan, Iran
mohsena@iaumajlesi.ac.ir
[2] Gwangju Institute of Science and Technology (GIST),
1 Oryong-dong Buk-gu, Gwangju, 500-712, Korea
hoyo@gist.ac.kr

Abstract. In this paper, we propose a methodology for hiding a gray scale guest image of low resolution into another gray scale host image of higher resolution. The embedding scheme can be used for secure or economic transmission of the low resolution image through the established communication channel for the high resolution image. We encode the guest image by a two-description subband coder. The information of one description is embedded in one frequency subband of the host image, and the other description in the other frequency subbands. We embed the information in those areas of the host image that contains high texture to reduce visibility of the embedded information in the host image. At the receiver, a multiple description decoder combines the information of each description to reconstruct the original guest image. We experiment the proposed scheme by embedding a gray scale guest image of 128*128 pixels in the gray-scale host image of 512*512 pixels, and evaluate the system robustness to various signal processing operations.

1 Introduction

Embedding images into other images and videos has applications in data hiding and digital watermarking. During the last few years, much progress has been made in developing watermarking techniques that are robust to signal processing operations, such as compression [1].

The signature information ranges from pseudo-random sequences to small image icons. In digital watermarking applications, emphasis is put on authentication rather than quantity and quality of the recovered data. It is necessary and satisfactory for the watermarking scheme to be able to prove the ownership, even though the host signal has undergone various signal processing or geometrical attacks. Data hiding has some other types of applications, such as broadcasting. In this application, the goal is to use an already established multimedia transmission channel for transmission of another multimedia signal. In this case, we need to recover the embedded information with high quality. However, in these applications the possibility of active and sever attacks is low. The host multimedia data could only be changed subject to some signal processing operations, such as compression, addition of noise, and down-sampling. As in

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 747–754, 2004.

digital watermarking, two scenarios are possible at the receiver: (a) referenced detection, where the original host signal is available, and (b) blind detection, where the original host signal is not available. The blind detection has a wider range of applications; however, it is more difficult to implement since the effect of the host signal on the embedded information cannot easily be removed at the receiver [1].

There have been few reports on large capacity data embedding [2],[3],[4] , Chae and Manjunath used the discrete wavelet transform (DWT) and lattice code for embedding a signature image/video into another image/video [2]. They further improved their system by using joint source-channel quantizers [3]. However; the channel-optimized quantizer is not suitable in data hiding applications, where intentional or non-intentional manipulations are variable and not known in advance.

In another approach Swanson et. al. [4] designed a method for embedding video in video based on linear projection in the spatial domain. The method is flexile and simple to implement, but like other spatial domain embedding techniques, it is not robust to compression [4].

In our previous work [5], we proposed an image data hiding scheme in the spatial domain. The system uses multiple description coding (MDC) of the guest image and informed data embedding in the spatial domain. In this paper, we propose a new data embedding and recovering scheme that does not need the host image at the receiver. We also use data embedding in the wavelet domain to reduce the visibility distortion. We embed the two descriptions in two independent channels to enjoy more from the potential of multiple description coding.

In case that the host signal is faced signal processing operations, we can retrieve a less corrupted description from the host image and reconstruct the signature image of acceptable quality using the less corrupted description.

After we provide an overview of the proposed image hiding scheme in Section 2, we explain the encoding process of the guest image using multiple description coding in Section 3. Section 4 explains the data embedding and extraction processes, and Section 5 explains the experimental results of the proposed scheme, and finally Section 6 summarizes the paper.

2 Overview of the Proposed Scheme

Fig. 1 shows the overall structure of the proposed scheme for guest image data embedding. We encode the guest image by a two-description subband coder. The two descriptions are represented by D_1 and D_2. The two portions of the host image are analogous to two communication channels for transmission of the two descriptors. The bitstream of the two descriptions are embedded in the two subbands.

At the receiver, we use only the received host image to recover the two descriptions, and reconstruct the signature image using the multiple description decoder.

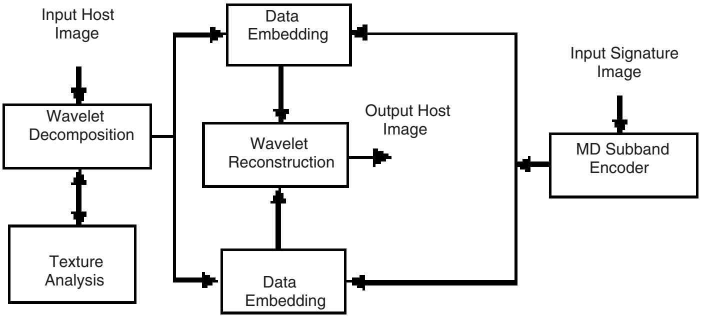


Fig. 1. Signature image embedding in the host image

3 Multiple Description Coding of the Guest Image

Multiple description coding was originally proposed for speech transmission over a noisy channel [6]. In this paper, we follow a simple but efficient approach of increasing redundancies in important portions of the image. Fig. 2 shows the overall structure of the signature image encoding operation. In the first stage, we decompose the signature image using the Haar wavelet transform, resulting in four subbands, usually referred to as LL, LH, HL, and HH. The lowest frequency subband (LL) contains the most important visual information and it needs more protection compared to high frequency subbands. As Fig. 2 shows, we use MD scalar quantizer for LL [7]. For high frequency subbands, we just split the pixels between the two descriptions.

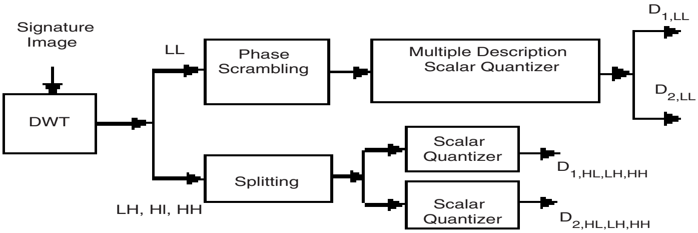


Fig. 2. Signature image encoding

Except for the lowest frequency subband (LL), the probability density function (PDF) for other subbands can be closely approximated with the Laplacian distribution. The LL does not follow any fixed PDF, but it contains the most important visual information. We use a phase scrambling operation to change the PDF of this band to a nearly Gaussian shape [8]. The added random phase could be an additional secret key between the transmitter and the registered receiver. We encode the two descriptions of the lowest frequency subband and the high frequency subbands with four state-trellis scalar quantizers. The average encoding bit-rate was 3 bit per sample and we obtained PSNR value over 38 dB for various test images, which is satisfactory in image hiding applications [2].

4 Data Embedding and Extraction in the DWT Domain

For our experiments, we select a gray-scale host image of 512× 512 pixels and guest image of 128 × 128 pixels size. We embed the information in the host image area with high texture content to be less visible. In order to evaluate the texture content of a block, we apply single-level Haar wavelet decomposition on each of the two portions of the host image [9]. As depicted in Fig. 3, each block k with 8 × 8 pixel size in the image is mapped to 4 blocks with 4 × 4 pixel size in the wavelet domain. We define a normalized measure for the energy of high frequency bands by

$$\mu_k = \left| \frac{e_H}{e_L} \right| \tag{1}$$

where e_H is the average of the absolute value of the high frequency bands (LH, HL, HH), and e_L is the absolute value of the lowest frequency band (LL) of the corresponding block. μ_k characterizes the given block texture energy. Higher value of shows the block has strong high frequency component or high texture. We consider these blocks good candidate for data embedding and replace some DCT or wavelet coefficients with embed data. In the following sections, we explain in more details the process of embedding in each domain.

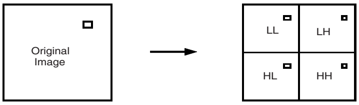


Fig. 3. The wavelet decomposition of the host image for texture measurement

We split the candidate blocks of the host image into two groups, and embed one description of the guest image in LH and the other in the HL band. For data embedding, we replace some wavelet coefficients in the selected blocks with the quantized value of the signature image description (D_1) after proper weighting by a defined modulation factor (α). Since the signature image size is lower than the host image size, we can select only part of the blocks using the texture measurement criterion derived in previous section, and replace some pixels of the total 64 pixels of the wavelet coefficients of a block with signature image quantized values.

At the receiver, we first reconstruct the lowest frequency subband of the guest image from the extracted indices in each portion of the host image, and recombine the two reconstructed bands. As the lowest frequency band is a blurred version of the original image. We can estimate the corrupted pixels from the over-sampled guest image using various error detection and concealment methods. In developed system, we follow a simple scheme based on comparison of each pixel with its neighboring pixel average value. Each pixel has four neighboring pixels from its own description and four neighboring pixels from the other description.

For the pixel with intensity value $x_{i,j}$ first we calculate the average value of neighboring pixels in the first descriptions: m_1 , and the second description: m_2; and then we calculate,

$$\lambda_1 = \mid \frac{x_{i,j} - m_1}{m_1} \mid \tag{2}$$

$$\lambda_2 = \mid \frac{x_{i,j} - m_2}{m_2} \mid \tag{3}$$

High value of λ_1 or λ_2 suggests possibility of corruption of the pixel. In this case we can replace the pixel $x_{i,j}$ with m_1 or m_2 . If the receiver can gain some knowledge of the type of attack the host image undergone, it can estimate which description has been more corrupted and adjust the or more efficiently.

For high frequency subbands, we can simply recombine the two independent descriptions. On the other hand, if based on evaluation from the lowest frequency subband we estimate that one of the two descriptions is highly corrupted, we can replace that description with the less corrupted description.

5 Experimental Results and Analysis

For our experiments, two images, ‘F16’ and ‘Einstein’, are used as signature images, and the ‘Shipping Boat’ image is used as the host images.

In order to control distortion resulted form data embedding in the host image; we can change the embedding factor in the wavelet domain. We set the modulation factor such that the host images PSNR stays above 38.5 dB for our further experiments. Fig. 4 shows samples of signature image recovered from the host images (‘Shipping Boat’) without any attacks.

As the goal of developed system has been image hiding for broadcasting application, we evaluate the system performance by calculating PSNR values of the reconstructed guest images. The system can be applied to applications such as hiding logo images for copyright protection, where the presence or absence of the signature is more important than the quality of the reconstructed image. In these applications, we usually set a threshold to decide on the amount of the cross-correlation between the recovered signature and the original signature [2]. However; in this paper, we concentrate only on image hiding applications and provide the reconstructed images PSNR values after the host image undergone some signal processing and geometric operations that could be the result of transmission or format exchange.

Fig. 4. Samples of reconstructed signature images

Resistance to Baseline-JPEG Compression: The JPEG lossy compression algorithm with different quality factors (Q) is tested. Fig. 5 shows the PSNR variation for different Q factors. As shown in Fig. 5, the PSNR values drop sharply for Q smaller than 50.

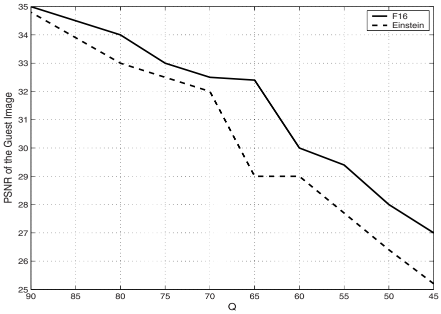


Fig. 5. PSNR of recovered signature images due to Baseline-JPEG compression

Resistance to Median and Gaussian Filtering: Median and Gaussian filters of 3 × 3 mask size are implemented on the host image after embedding the signature. The PSNR of recovered signature are shown in Table 1.

Table 1. PSNR (dB) values of the recovered signature images after implementing median and Gaussian filters on the host image

	Median Filter	Gaussian Filter
F16	26.28	30.23
Einstein	25.64	28.52

Resistance to Cropping: In our experiments we cropped parts of the host image coroners. Fig. 6 shows sample of host image after 20% cropping. We filled the cropped area with the average value of remaining part of the image. Table 2 shows PSNR values when some parts of the host image corners are cropped. Considerably good resistance is due to the existence of two descriptors in the image and scrambling of embedded information, which makes it possible to reconstruct the signature image information partly in the cropped area from the available descriptor in the non-cropped area.

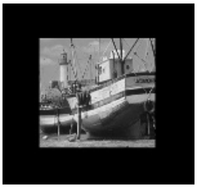


Fig. 6. Sample of the host image after 20% cropping

Table 2. PSNR (dB) values of the recovered signature image for different percentage of cropping the host image

	5%	10%	15%	20%
F16	26.78	25.83	22.40	21.65
Einstein	25.26	24.44	23.30	20.01

Resistance to JPEG2000 Compression: The JPEG2000 lossy compression algorithm with different output bit rates is tested on the host image. Fig.7 shows the PSNR variation of the recovered signature images.

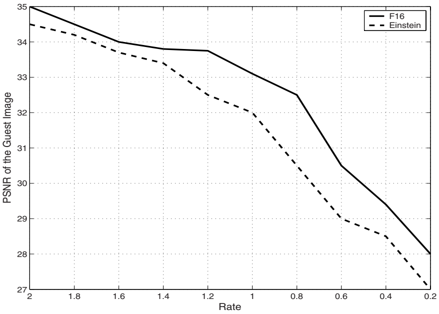


Fig. 7. PSNR variation of recovered signature images due to JPEG-2000 compression

6 Conclusions

We have presented a new scheme for embedding a gray scale image into another gray scale image. We developed an MDC scheme for encoding guest image based on MD scalar quantization of the lowest frequency subband and splitting the high frequency bands. The encoded bitstream of the two descriptions was embedded in different frequency subbands. We used a measure for evaluating texture content of the host image blocks for data embedding to reduce the visibility distortion. Results of various experiments on recovering the guest image after the host image is undergone different attacks, show the MDC of the guest image and embedding in different frequency subbands make it possible to recover the guest image with good quality. The embedding scheme can be further improved by having a prior knowledge or feedback about types of attacks or processing that host image might have undergone. The system can be extended for hiding a video signal in another video signal for transmission.

Acknowledgments. This work is supported by Islamic Azad University of Iran. Also it was supported in part by GIST, in part by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC) at GIST, and in part by the Ministry of Education (MOE) through the BK21 project.

References

1. Petitcolas, F.A.P., Anderson R.J., and Kuhn, M.G.: Information Hiding-a Survey, Proceedings of the IEEE, Vol.87, No.7, (**1999**) 1062–1078
2. Chae J.J and Manjunath, B.S.:A Robust Embedded Data from Wavelet Coefficients. in Proceeding of SPIE: Storage and Retrieval for Image and Video Databases, Vol. IV, Jan. (**1998**) 308–317
3. Mukherjee, J.J., et. al.: A Source and Channel-Coding Framework for Vector-Based Data Hiding in Video. IEEE Trans. on Circuits and System for Video Technology, Vol. 10, No.6, (**2000**) 630–645
4. Swanson, M.D., et. al.: Data Hiding for Video-in-Video. IEEE International Conference on Image Processing, Vol. 2, (**1997**) 676–679
5. Ashourian, M., Ho, Y.S.: Multiple Description Coding for Image Data Hiding in the Spatial Domain. Lecture Notes in Computer Science (LNCS), Vol. 2869, (**2003**) 659–666
6. Jayant, N.S.: Sub-sampling of a DPCM Speech Channel to Provide Two Self-contained Half-rate Channels. Bell System Technical Journal, Vol.60, No.4, (**1981**) 501–509
7. Vaishampayan, V.A.: Design of Multiple Description Scalar Quantizers. IEEE Trans. on Information Theory, Vol. 39, No.5, (**1993**) 821–834
8. Kuo C.C.J. ,Hung, C.H.: Robust Coding Technique-Transform Encryption Coding for Noisy Communications. Optical Eng., Vol. 32, No. 1, (**1993**) 150–153
9. Chae J.J. Manjunath, M.J. :A Technique for Image Data Hiding and Reconstruction without Host Image. in Proc. of SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, Vol. 3657, San Jose, California, (**1999**) 386–396

Image Watermarking Capacity Analysis Using Hopfield Neural Network

Fan Zhang and Hongbin Zhang

Institute of Computer, Beijing University of Technology, Beijing 100022, China
bpuzf@sohu.com

Abstract. Image watermarking capacity research is to study how much information can be hidden in an image. In watermarking schemes, watermarking can be viewed as a form of communication and image can be considered as a communication channel to transmit messages. Almost all previous works on watermarking capacity are based on information theory, using Shannon formula to calculate the capacity of watermarking. This paper presents a blind watermarking algorithm using Hopfield neural network, and analyze watermarking capacity based on neural network. Result shows that the attraction basin of associative memory decides watermarking capacity.

1 Introduction

Capacity is a very important property of digital watermarking. The purpose of watermarking capacity research is to analyze the limit of watermarking information while satisfying watermarking invisibility and robustness.

Several works on watermarking capacity have been presented in recent years. Servetto considered each pixel as an independent channel and calculated the capacity based on the theory of Parallel Gaussian Channels (PGC) [1]. Barni's research focused on the image watermarking capacity in DCT and the DFT domain [2]. Moulin's work studied a kind of watermarking capacity problem under attacks [3,4]. Lin presented zero-error information hiding capacity analysis method in JPEG compressed domain using adjacency-reducing mapping technique [5,6].

In watermarking schemes, watermarking can be considered as a form of communications. Image is the communication channel to transmit messages, and the watermark is the message to be transmitted. So, watermarking capacity problem can be solved using traditional information theory. Almost all previous works on image watermarking capacity are based on information theory, using Shannon formula to calculate the capacity of watermarking. In this paper, we present a blind digital watermarking algorithm using Hopfield neural network, and analyze watermarking capacity based on neural network.

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 755–762, 2004.

2 Blind Watermarking Algorithm Based on Hopfield Neural Network

In this section, we present a blind watermarking algorithm based on Hopfield neural network, using the Noise Visibility Function (NVF) [7,8] for adaptive watermark embedding. Our idea is to store host (original) image and original watermark using associative Hopfield neural network, and to retrieve host image and original watermark when watermark extracting.

2.1 Adaptive Watermarking

Noise Visibility Function is the function that characterizes local image properties, identifying textured and edge regions where the watermark should be more strongly embedded.

Assuming the host image subject to generalized Gaussian distribution, the NVF at each pixel position can be written as:

$$NVF = \frac{1}{1 + \sigma_x^2(i,j)}, \tag{1}$$

where σ_x^2 is the local variance of the image in a window centered on the pixel with coordinates (i,j). Once we have computed the NVF, we can obtain the allowable distortions of each pixel by computing:

$$\Delta(i,j) = (1 - NVF(i,j)) \cdot S_0 + NVF(i,j) \cdot S_1, \tag{2}$$

where S_0 and S_1 are the maximum allowable pixel distortions in texture and flat regions respectively. Typically S_0 may be as high as 30 while S_1 are usually about 3. In flat regions the NVF tends to 1, so the first term of Eq.(2) tends to 0, and consequently the allowable pixel distortion dependent on S_1. Intuitively this makes sense, since we expect that the watermark distortions will be more visible in flat regions and less visible in texture regions. According to above equation, the watermark embedded in texture or edge regions is stronger than in flat regions. If we embed maximum allowable watermark in each pixel, the robustness of watermarking will have a good performance. By this way, we can achieve the best trade-off between robustness and invisibility.

2.2 Blind Watermarking Using Neural Network

Hopfield neural network is a nonlinear dynamical system, using computational energy function to evaluate the stability property. The energy function always decreases toward a state of the lowest energy. Starting from any point of state space, system will always evolves to a stable state, an attractor.

Neurons state of Hopfield neural network are usually binary {+1, -1}. For the sake of neural network to store a standard grayscale test image, we decompose image into eight bit planes. For each pixel, we decide whether the pixel should be modified randomly by using a random function. 0 denotes do not modify this pixel and 1 denotes modify this pixel. Then we can get a matrix that composes of (0,1). We name the matrix Watermark Bit Plane (WBP). Actually, the WBP mark the place of watermark embedding in image. Watermark amplitude is decided by Eq.(2). During the learning of neural network, not only image bit planes but also Watermark Bit Plane are stored by Hopfield neural network. In watermark extracting, network recalls the host (original) image and WBP, and then we compare the retrieved host image with stego (watermarked) image, extract watermark according a threshold. Finally, comparing retrieved WBP with extracted watermark, we can judge whether watermark exist in the image using correlation test.

3 Watermarking Capacity Analysis Based on Neural Network

Hopfield neural network can work as an associative memory, the patterns are stored as dynamical attractors, and the network has error-correcting capability. Basin of attraction is the set of points in the space of system. The radius of attraction basin is defined as the largest Hamming distance within which almost all states flow to the pattern. For a trained network, the average attraction radiuses of stored patterns gives a measure of the network completion capability.

The Hamming distance of two vectors S^1 and S^2 is the number of components different from each other. We mark it as $d_h(S^1, S^2)$. The total number of vectors which Hamming distance to S^p less than r constitutes the r-Hamming sphere:

$$B_r(S^p) = (S^q|d_h(S^p, S^q) = r). \tag{3}$$

Basin of attraction can be denoted by Hamming distance or Hamming sphere, which represents the error-correcting capability of neural network.

Assume P denotes the number of stored patterns and N denotes the number of neurons. Hopfield associative memory model can be expressed as:

$$x_i^{t+1} = sgn\left[\sum_{j\neq i}^{N} W_{ij} x_j^t\right], \tag{4}$$

where x_i^t denotes neuron's state at time t; sgn is the sign function. The connection weight matrix can be computed according to Hebb rule:

$$W_{ij} = \frac{1}{N}\sum_{k=1}^{P} u_i^k u_j^k, \tag{5}$$

where u^1, u^2,…, u^P denotes stored patterns of neural network. Assume that $X^0 = \{x_1^0, x_2^0, \ldots, x_N^0\}^T$ denotes neural network's initial state, $X^t =$

$\{x_1^t, x_2^t, \ldots, x_N^t\}^T$ denotes neural network's state at time t. If we assume the probe pattern is one of stored patterns, according to Eq.(4) and (5), then:

$$\sum_{j \neq i}^{N} W_{ij} x_j^0 = \sum_{j \neq i}^{N} \frac{1}{N} \sum_{k=1}^{P} u_i^k u_j^k x_j^0$$

$$= \frac{1}{N} \sum_{k=1}^{P} \left[\left(\sum_{j=1}^{N} u_i^k u_j^k x_j^0 \right) - u_i^k u_j^k x_i^0 \right]$$

$$= \frac{1}{N} \left[u_j^k (u^k)^T X^0 + \sum_{l \neq k}^{P} u_i^l (u^l)^T X^0 - P x_i^0 \right]. \tag{6}$$

Because $(u^k)^T X^0 = N - 2d_h(X^0, u^k)$ and $\frac{N-P}{2P} < \frac{N}{2}$, If we assume x_i are orthogonal for each other, according to Eq.(20) of Appendix, then:

$$-2d_h(X^0, u^k) \leq (u^l)^T X^0 \leq 2d_h(X^0, u^k). \tag{7}$$

When $d_h(X^0, u^k) < \frac{N-P}{2P}$, then

$$N - 2d_h(X^0, u^k) - 2(P-1)d_h(X^0, u^k) - P > 0. \tag{8}$$

So, when $u_i^k = +1$,

$$\sum_{j \neq i}^{N} W_{ij} x_j^0 = \frac{1}{N} \left[u_j^k (u^k)^T X^0 + \sum_{l \neq k}^{P} u_i^l (u^l)^T X^0 - P x_i^0 \right]$$

$$> \frac{1}{N} \left[N - 2d_h(X^0, u^k) - 2(P-1)d_h(X^0, u^k) - P \right] > 0. \tag{9}$$

when $u_i^k = -1$,

$$\sum_{j \neq i}^{N} W_{ij} x_j^0 < \frac{1}{N} \left[-N + 2d_h(X^0, u^k) + 2(P-1)d_h(X^0, u^k) + P \right] < 0. \tag{10}$$

So, $u_i^k = sgn \left[\sum_{j \neq i}^{N} W_{ij} x_j^0 \right]$, and then:

$$x_i^1 = sgn \left[\sum_{j \neq i}^{N} W_{ij} x_j^0 \right] = u_i^k. \tag{11}$$

According to Eq.(11), if the Hamming distance between the probe pattern and stored patterns,

$$d_h \leq \frac{N-P}{2P}, \tag{12}$$

the Hopfield neural network can associative recall the stored pattern at the first time.

In watermarking schemes, watermark can be viewed as noise that pollutes the original image. If we modify amplitude of some pixels, then there will occur some change in corresponding place on bit planes. This means that those bit planes are polluted. The more watermark embed, the bigger Hamming distance between stego image with host image is. When the Hamming distance is out of the bound of attraction basin, neural network cannot recall host image correctly. So, the bound of attraction basin confines the number of points in the image that can be modified, furthermore, confines the capacity of watermarking.

Equation (12) is derived when stored patterns are orthogonal. Obviously, if stored patterns are nonorthogonal, basin of attraction is less than d_h. For analyzing the limit of information, we assume stored patterns are orthogonal. We think it is reasonable, it just like we assume information source and noise subject to Gaussian distribution during channel capacity analysis in information theory.

According to our watermarking algorithm, there are nine stored patterns. According to Eq.(12), $d_h \leq 3640$. If the number of modified points of a bit planes less than d_h, neural network can associative recall bit planes successfully.

Modifying a pixel may results in several bit plane's change in corresponding place. If a bit plane be modified, corresponding pixel is surely be modified. Contrarily, if the place of a bit plane is not modified, we cannot affirm this pixel is not embedded watermark, maybe other bit planes of this place are modified.

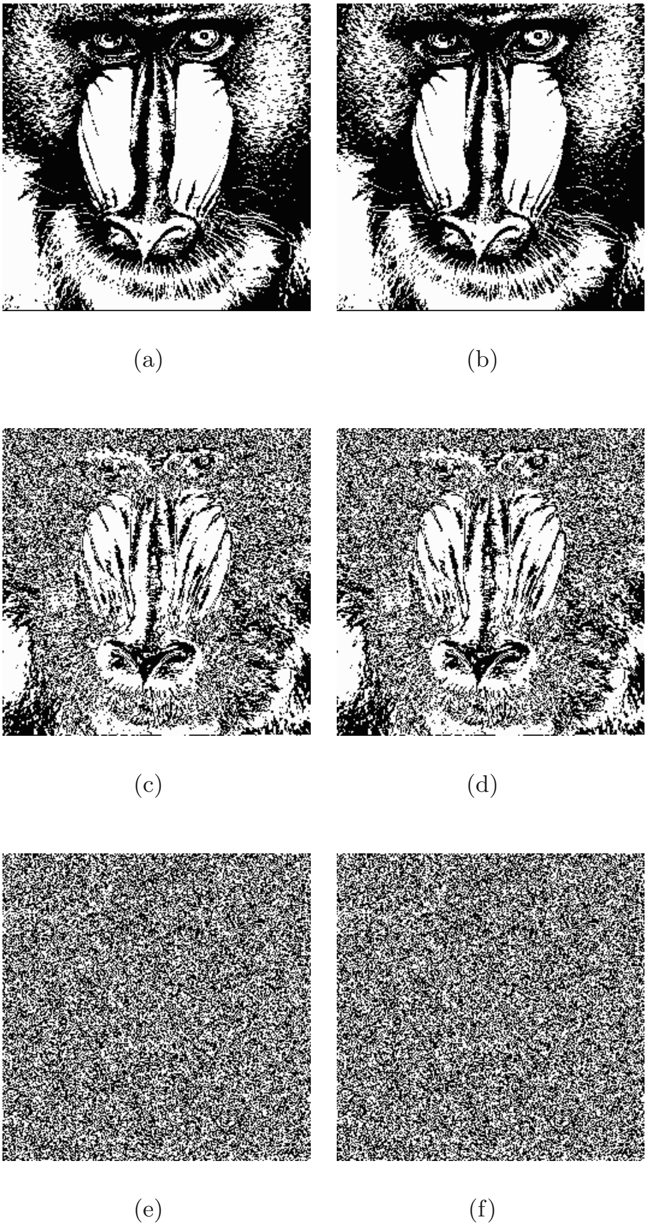
Because the maximum watermark amplitude is no more than 30 (decided by S_0 in Eq.(2)), when embedding watermark in a pixel, at least one of five low bit planes is modified in corresponding point. As the analysis above, the maximum number of modifiable points is 3640. In an extreme case, modified points in five low bit planes may different to each other, then the maximum number of watermarking pixels is $n = 5 \times 3640 = 18200$. In our watermarking algorithm, watermark is a binary sequence. State of each component in the sequence is random (decided by a random function). A n-length binary sequence has $2n$ combinations in all, each combination appear in probability $1/2n$. According to information theory, information of an n-length binary sequence is:

$$C = -log_2(\frac{1}{2^n}). \tag{13}$$

So, in our watermarking algorithm, watermarking capacity is 18200 bits.

4 Experiments

In experiments, three 256 × 256 standard test images Baboon, Peppers and Lena are used. If the number of neurons equals to the number of pixels, the computation will be very complex. So we should reduce the dimension of neural network to reduce the commutative complexity. The NVF is calculated in a local image region, a window centered on a pixel. We divide image into many non-overlap



(a) (b)

(c) (d)

(e) (f)

Fig. 1. Original Baboon image bit planes (a), (c), (e) and retrieved image bit planes (b), (d), (f).

Table 1. Positive detection (presence of watermark) times of one hundred times experiments per test images in different conditions.

	Baboon	Peppers	Lena
stego image	100	100	100
noised stego image	100	100	100
host image	0	0	0
noised host image	0	0	1

regions according to the window size of the NVF. Each region corresponds to a neuron. If the NVF calculation region is a window of size 5×5, then a 256×256 image can be divided into 51×51 regions, and the dimension of neural network is 51×51.

In watermark extracting, we assume that the probe patterns are stego image, noised stego image, host image and noised host image respectively. We experimented 1200 times, one hundred times in above conditions for each test image. Table (1) shows the result of our experiments. Figure (1) shows original Baboon image bit planes and retrieved image bit planes.

5 Conclusions

Almost all previous works on image watermarking capacity are based on information theory, using Shannon formula to calculate the capacity of watermarking. This paper presents a blind watermarking algorithm using Hopfield neural network, using the NVF for adaptive watermark embedding. And we analyze watermarking capacity based on neural network. Result shows that the attraction basin of associative memory decides watermarking capacity.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (60075002) and the National High Technology Research and Development 863 Program of China (2001AA144080).

References

1. S D Servetto, C I Podilchuk, K Ramchandran. Capacity Issues in Digital Image Watermarking. IEEE International Conference on Image Processing. 1998, vol.1:445–449
2. M Barni, F Bartolini, A De Rosa, A Piva. Capacity of the watermarking-channel: how many bits can be hidden within a digital image. Security and Watermarking of Multimedia Contents, Proceedings of SPIE. 1999, Vol.3657:437–448
3. P Moulin, M K Mihcak. A Framework for Evaluating the Data-Hiding Capacity of Image Sources. IEEE Transactions on Image Processing. 2002,Vol.11, No. 9:1029–1042

4. P Moulin. The Role of Information Theory in Watermarking and Its Application to Image Watermarking. Signal Processing. 2001,Vol.81, No.6:1121–1139
5. C Y Lin, S F Chang. Zero-error Information Hiding Capacity of Digital Images. IEEE International Conference on Image Processing. 2001,Vol.3:1007–1010
6. C Y Lin. Watermarking and Digital Signature Techniques for Multimedia Authentication and Copyright Protection. Ph.D. Thesis. Columbia University, 2000
7. S Voloshynovskiy, S Pereira, V Iquise, T Pun. Towards a second-generation benchmark. Signal Processing. 2001, Vol.81, No. 6:1177–1214
8. S Pereira, S Voloshynovskiy, T Pun. Optimal transform domain watermark embedding via linear programming. Signal Processing. 2001, Vol.81, No.6: 1117–1135

Appendix

For any N-length vectors P, Q, X,

$$\begin{aligned} P^T Q &= N - 2d_h(P, Q), \\ X^T Q &= N - 2d_h(X, Q), \\ X^T Q &= N - 2d_h(X, Q). \end{aligned} \tag{14}$$

Using triangle distance inequation:

$$d_h(P, Q) + d_h(X, P) \geq d_h(X, Q), \tag{15}$$

we can deduce:

$$\frac{N - P^T Q}{2} + \frac{N - X^T P}{2} \geq \frac{N - X^T Q}{2}, \tag{16}$$

and

$$\begin{aligned} X^T Q &\geq P^T Q + X^T P - N, \\ X^T Q &\geq -P^T Q - X^T P - N. \end{aligned} \tag{17}$$

According to Eq.(17), then:

$$\left|P^T Q + X^T P\right| - N \leq X^T Q. \tag{18}$$

Similarly, we can deduce:

$$X^T Q \geq N - \left|P^T Q - X^T P\right|. \tag{19}$$

If P, Q are orthogonal, $P^T Q = 0$, then:

$$\left|X^T P\right| - N \leq X^T Q \leq N - \left|X^T P\right|. \tag{20}$$

Efficient Algorithms in Determining JPEG-Effective Watermark Coefficients

Chih-Wei Tang and Hsueh-Ming Hang

Department of Electronics Engineering,
National Chiao Tung University,
Hsinchu 30050, Taiwan.
chihwei.ee88g@nctu.edu.tw, hmhang@mail.nctu.edu.tw

Abstract. A set of coefficient selection rules is proposed for efficiently determining the effective DCT watermarking coefficients of an image for JPEG attack. These rules are simple in computation but they are derived from the theoretically optimized data set with the aid of the parametric classifiers. They improve the watermark robustness (correctly decoding) and, in the mean time, decrease the error detection probability (correct detection). The frequency versus watermark strength space is used in constructing the selection rules. Simulation results show that the computational complexity is significantly reduced compared to our previous theory-based optimization work, but still the selected coefficients can achieve nearly the same performance as the original scheme.

1 Introduction

Many digital watermarking schemes have been proposed for copyright protection, data hiding and other purposes. In our previous work, we focus on the tradeoffs between the achievable watermarking data payload, allowable distortion for information hiding, and robustness against attacks [1]. Although many methods have been developed to improve the watermark data payload and robustness while maintaining reliable detection and visual fidelity [2]-[5], few researchers have proposed techniques to identify the exact coefficient locations for watermarking. Thus, we suggested a generic approach for selecting the most effective coefficients for watermark embedding. Using this set of coefficients improves the watermark robustness and reliability while it maintains the watermark visual transparency. To a certain extent, we try to find the performance limit of invisible watermarking for a given natural image under the assumptions of known attack and non-blind detection for DCT-domain watermarking. The non-blind detection can be used in applications such as transaction-tracking. The synchronization attack is not considered as a problem due to non-blind detection. Since digital images are often compressed for efficient storage and transmission, we use JPEG and JPEG2000 as the examples of attacking sources in the design phase.

Although the coefficient selection procedure performs rather well, its computational complexity is very high. Therefore, in this paper, we develop a fast algorithm with nearly no performance loss. Due to the limited space, only the

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 763–770, 2004.

simplified rules for JPEG compression attacking source is presented. Note that the methodology of the coefficient selection procedure in [1] and the simplified algorithm proposed in this paper both can easily be extended to the other types of attacks. Section 2 briefly describes our previous work theory-based optimal coefficient selection. Section 3 describes the newly proposed coefficient selection rules. Simulation results are summarized in Section 4 and Section 5 concludes this presentation.

2 Our Previous Optimization Algorithm

Two optimization stages are proposed in [1] for selecting effective coefficients. One is the robust and imperceptible coefficient selection stage (Stage One), and the other is the detection reliability improvement stage (Stage Two). Stage One conducts a deterministic analysis on the transform coefficients, and then the proper coefficients and the associated watermark strength are determined so that the coefficients after a specified attack can still bear the valid marks. The additive embedding is adopted in the DCT domain, where is the watermark strength of the ith AC coefficient $x[i]$ and $w[i]$ is the watermark bit. All AC coefficients are watermarked. For an attack in either the spatial or other transform domains, the watermarked image is converted back to the spatial domain and the attack is applied. We decode the watermark bits in the DCT-domain. Several different watermark patterns are tested. If all watermark bits associated with a certain DCT coefficient are correctly decoded, this coefficient is retained in the Stage One candidate set. We examine the all-positive and all-negative watermark patterns. When the attack is not applied to individual coefficients in the DCT-domain, we also test the alternate polarity pattern in which the odd-index watermark bits (in zigzag scan order) are +1 and the even-index ones are -1. This is because the attack distortion on a DCT coefficient also depends on its neighboring watermark bits. Our experiments indicated that we can identify robust coefficients with rather high probability by only 4 patterns. The Watson's visual model is adopted for contrast masking threshold computation and the parameter values are taken from the Checkmark package [6].

Some robust coefficients may produce higher detection error probability. Thus, Stage Two calculates the statistical measures on images and attacks, and it discards the weak coefficients. An iterative procedure is proposed and only one coefficient is discarded in each iteration. At the beginning of one iteration, if N coefficients remain, N candidate sets are formed by deleting one coefficient alternatively in this N-coefficient set. That is, there are N-1 coefficients in each candidate set. Then, the watermark detection statistics based on signal dependent channel distortion model [7] and the Bayes'decision rule for each candidate set is calculated for each candidate set. The error detection probability is the average of the false positive probability and false negative probability. Then, the set with the lowest detection error probability is chosen if the average error probability decreases from the previous iteration. The coefficient discarding process is repeated until the overall error probability cannot be further reduced. If there are N selected coefficients at the beginning of Stage Two, and K dropped coefficients in the process, the execution time of Stage Two will be $O(KN^2)$. Thus, a fast algorithm is very desirable.

3 Efficient Robust and Reliable Coefficient Selection Rules

Our goal is finding simplified rules to separate the selected coefficients and dropped coefficients for a given input image based on the theoretically optimized data set derived from [1]. We adopt a parametric linear classifier for classification [8]. For a parametric approach, most of the estimated expressions are functions of expected vectors and covariance matrices. Although linear classifiers are not optimum, we use it due to its simplicity. The classifier (linear discriminate function) is

$$h(X) = V^T X + v_0, \tag{1}$$

where X is the given input data vector which distributions are not limited, $V = [v_1 v_2 \ldots]^T$ is the coefficient vector, and v_0 is a threshold value. To find the optimal V^T and v_0 for a given distribution, the criterion g is maximized, which measures the between-class scatter normalized by the within-class scatter,

$$g = \frac{P_1 \eta_1^2 + P_2 \eta_2^2}{P_1 \sigma_1^2 + P_2 \sigma_2^2}, \tag{2}$$

where P_i, η_i, and σ_i are the priori probability, expected value of $h(X)$, and variance of $h(X)$ for class i, respectively. As a result,

$$V = [P_1 \Sigma_1 + P_2 \Sigma_2]^{-1} (M_2 - M_1), \tag{3}$$

$$v_0 = -V^T [P_1 \Sigma_1 + P_2 \Sigma_2], \tag{4}$$

where Σ_i is the covariance matrix for a given expected vector X. Here, the well known fisher criterion is not adopted since it cannot determine the optimum v_0. The features in our problem are frequency f, amplitude x and admissible watermark strength α. Our target is to find a piece-wise linear classifier (discriminator) that separate the selected coefficients from the dropped ones. We have looked at the case that uses all three features (f,x,α) (3-D domain). To simplify calculations, we also search for a 2-D feature space with smallest average misclassification rate. Our experiments show that the "optimal" average misclassification rate in the 2-D space "(f,α)" is only 1% lower than that of the 3-D domain classifier. There are three 2-D domain candidates: (f,x), (f,α), and (x,α). Let D_{fx}, $D_{f\alpha}$ and $D_{x\alpha}$ be the misclassification rate due to the selected coefficients are misclassified as dropped coefficients in the aforementioned three candidate spaces, respectively, S_{fx} and $S_{f\alpha}$, and $S_{x\alpha}$ be the misclassification rate due to the dropped coefficients are misclassified as selected coefficients. To decrease S_{fx} and $S_{f\alpha}$, and $S_{x\alpha}$, we set $P_1 = 0.4$ and $P_2 = 0.6$. For further improving the classification accuracy, we divide a space into three subspaces, and design one linear classifier for each subspace. For (f,x) and (f,α) spaces, the separation is based on f=0-9, f=10-19 and f=20-63. For space (x,α), they are x=0-49, x=50-99 and x=100-∞. Our image data base contains 30 natural images. The training set is generated using the method described in Sect. 2. Four JPEG quality factors ranging from 50 to 80 are used. We adopt the definition of JPEG quantization step size defined in [9]. The misclassification rates in all cases (2D domains) are listed in Table 1. Because the best 2-D (f,α) space is 1% worse than the 3-D (f,x,α) classifier, the former is adopted for a much lower computation complexity.

Table 1. Misclassification rates in three 2-D feature spaces

Design Phase	S_{fx}	$S_{f\alpha}$	$S_{x\alpha}$	D_{fx}	$D_{f\alpha}$	$D_{x\alpha}$
JPEG50	0.31	**0.18**	0.66	0.07	**0.10**	0.40
JPEG60	0.29	**0.18**	0.65	0.06	**0.09**	0.38
JPEG70	0.27	**0.18**	0.63	0.06	**0.09**	0.35
JPEG80	0.27	**0.16**	0.57	0.05	**0.08**	0.30

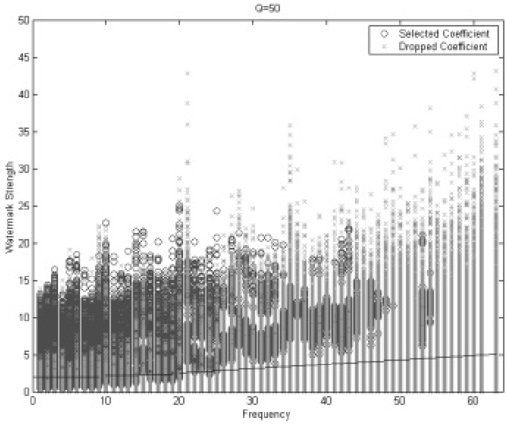


Fig. 1. The classifier at JPEG quality factor 50 with coefficients from 30 natural images

Thus, we can now select effective watermarking coefficients with the simplified rules. Figure 1 shows the classifier (coefficient selecting rules) for the JPEG quality factor 50 in the design phase. Although these rules eliminate a number of poor candidate coefficients, the remaining coefficients do not necessarily have the required robustness. Therefore, we apply the original Stage One process to the retained coefficients for further removing weak coefficients.

4 Simulation Results

To examine the performance of the proposed rules, we test images which are not used in training. Limited by space, only the results for pictures Lena and Baboon are included. For the JPEG quality factor 50 in the design phase, the PSNR values between the original and the watermarked images are 45.2 dB and 39.98 dB for Lena and Baboon, respectively. And, they are 42.9 dB and 36.82 dB for JPEG quality factor 80 in the design phase. The embedded watermarks are invisible as we inspect them visually. The comparisons between the original and the simplified schemes are shown in Tables 2 and 3. Let the overlapped percentage be the number of coefficients selected by both the original Stage One and the simplified scheme divided by the number of selected coefficients by the original Stage One. We find that the overlapped percentage is higher than 70%. The detection error probability using the simplified scheme is still very small (all less than 10^{-135} for Lena). Practically these rules are as good as the original massive iteration scheme. In the case of Baboon image, the overlapped percentage is over 85% and the detection error probability is all less than 10^{-245}. The data shown in Fig. 2 is each averaged over 5000 watermarked images with different random watermark sequences. Also, the same 5000 watermark sequences are correlated with the unmarked but JPEG compressed image and the results are averaged in Fig. 3. Figure 3 shows that the selected coefficient survives JPEG compression at higher quality factors may not survive JPEG compression at lower quality factors. To verify the designed false negative and positive error probabilities, the mean, variance, minimum and maximum values of the normalized correlation sum after the JPEG attacks are computed. (The normalization is normalized against the embedded watermark power as discussed in [1], and thus is not bounded to [-1, 1].) Due to the limited space, only the mean of the normalized correlation sum for watermarked images is shown in Fig. 2. The mean value of the normalized correlation sum C is computed by

$$C = \frac{1}{M}\sum_{i=1}^{M} c[i] = \frac{1}{M}\sum_{i=1}^{M} \frac{y[i]\times(w[i]\times\alpha[i])}{\sigma_d^2}, \qquad (5)$$

where $y[i]$ is the difference between the DCT coefficients of the received image and the original image, $w[i]$ is the watermark signature and M is the number of selected coefficients. For a watermark sequence, C is compared against the detection threshold which is approximately the average of the mean values of the normalized correlation sum of the watermarked $E\{c|H_1\}$ and unmarked images $E\{c|H_0\}$[1]. The presence of the watermark is declared if H_1 is favored. In all cases, there is no failure for either watermarked or unmarked 5000 images. Finally, small variance implies lower error detection probability. The variance values $Var\{c|H_0\}$ and $Var\{c|H_1\}$ are all smaller than 0.0018 after JPEG attacks with different quality factors for both watermarked and unmarked cases. We also test the JPEG-robust watermark against several other signal processing attacks by as shown in Fig. 4 and the data are obtained by averaging over 100 different random watermark sequences. The $E\{c|H_1\}$ is over 0.8 after JPEG2000

Table 2. The comparisons of the selected coefficients for Lena

Design Phase	No. of Selected Coeff. by Org. Stage 1	No. of Selected Coeff. by Org. Stage 2	Estimated P_{error} after Org. Stage 2	No. of Selected Coeff. by Fast Scheme	Estimated P_{error} by Fast Scheme
JPEG50	4738	4019	5.505e-299	3609	2.097e-136
JPEG60	6007	5082	0.000e+000	4516	6.803e-181
JPEG70	8041	6587	0.000e+000	5911	2.320e-253
JPEG80	111473	9439	0.000e+000	8166	0.000e+000

Table 3. The comparisons of the selected coefficients for Baboon

Design Phase	No. of Selected Coeff. by Org. Stage 1	No. of Selected Coeff. by Org. Stage 2	Estimated P_{error} after Org. Stage 2	No. of Selected Coeff. by Fast Scheme	Estimated P_{error} by Fast Scheme
JPEG50	9270	7359	0.000e+000	7877	3.099e-246
JPEG60	11743	8972	0.000e+000	10130	0.000e+000
JPEG70	15912	13105	0.000e+000	13708	0.000e+000
JPEG80	22931	18885	0.000e+000	20120	0.000e+000

Table 4. The comparisons of the selected coefficients for Baboon

Design Phase	No. of Processed Coeff. by Org. Stage 1	No. of Processed Coeff. by Org. Stage 2	No. of Processed Coeff. by Fast Scheme
JPEG50	64512	3152520	73629
JPEG60	64512	5134207	74108
JPEG70	64512	10641870	75139
JPEG80	64512	21277960	76366

attacks at bit rates 0.125 bpp and 0.0625 bpp. We also compare the computational complexity between the original and the simplified stages as shown in Table 4. The computational complexity is expressed by the number of processed DCT coefficients. For image Lena at JPEG quality factor 80, the simplified scheme requires roughly $\frac{1}{266}$ of the computations of the original scheme (Stage One + Stage Two) for large candidate sets. The simplified scheme does greatly reduce the computational complexity.

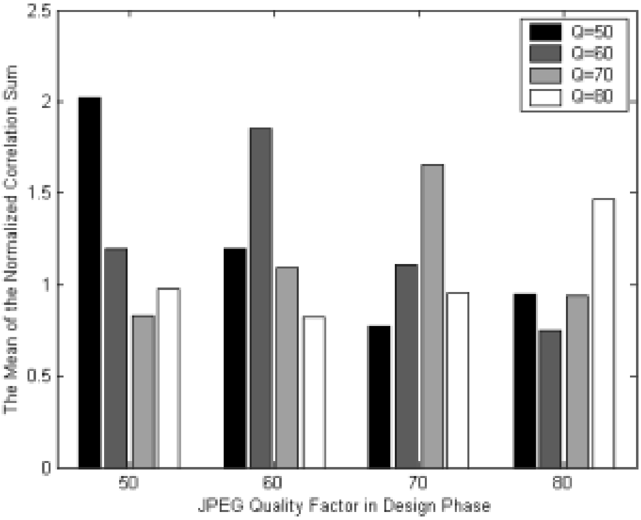


Fig. 2. The mean of the normalized correlation sum after JPEG attacks at different quality factors for watermarked Lena

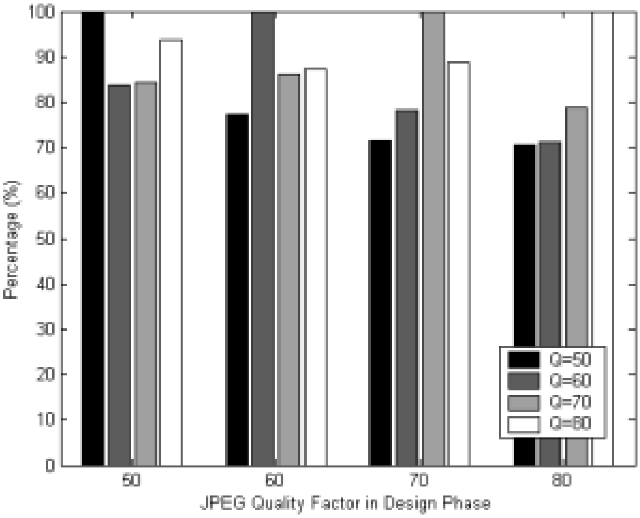


Fig. 3. The percentage of correctly decoded coefficients at the detector after JPEG attacks for Lena

5 Conclusions

In this paper, we propose an efficient algorithm for selecting JPEG-effective watermark coefficients. In most cases, the new scheme uses only $\frac{1}{100}$ of the computation needed in the original scheme in [2]. The methodology of both the original coefficient selection procedure in [2] and the simplified algorithm proposed here can be easily extended to the other types of attacks.

Acknowledgements. This work is partially supported by the Lee and MTI Center for Networking Research at National Chiao Tung University, Taiwan.

References

1. C.-W. Tang and H.-M. Hang: Exploring Effective Transform-Coefficients in Perceptual Watermarking. Proc. SPIE Security and Watermarking of Multimedia Contents IV. **4675** (2003) 572–583
2. P. Moulin and J. A. O'Sullivan: Information-Theoretic Analysis of Information Hiding. IEEE Trans. Information Theory. **49** (2003) 563–593
3. Q. Cheng, Y. Wang and T. S. Huang: How to Design Efficient Watermarks? IEEE International Conference on Acoustics, Speech, and Signal Processing. **3** (2003) 49–52
4. M. Barni, F. Bartolini, A. D. Rosa and A. Piva: Optimum Decoding and Detection of Multiplicative Watermarks. IEEE Trans. Signal Processing. **51** (2003) 1118–1123
5. A. Giannoula, A. Tefas, N. Nikolaidis and I. Pitas: Improving the Detection Reliability of Correlation-Based Watermarking Techniques. ICME. **1** (2003) 209–212
6. http://watermarking.unige.ch/Checkmark/index.html
7. J. J. Eggers and B. Girod: Quantization Effects on Digital Watermarks. Signal Processing. **831** (2000) 239–253
8. K. Fukunaga: Introduction to Statistical Pattern Recognition. Academic Press. (1990)
9. http://www.cl.cam.ac.uk/ fapp2/watermarking/stirmark/

A Fragile Watermarking Based on Knapsack Problem

Hui Han, HongXun Yao, ShaoHui Liu, and Yan Liu

School of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, P.R. China
{hh,yhx,shaohl,lyan}@vilab.hit.edu.cn

Abstract. In this paper, a new fragile watermarking is proposed. It is different from most existing schemes to resist the famous birthday attack. The classic NPC problem knapsack problem in algorithm is introduced. It uses a random sequence to encrypt the bit planes of image pixels and regard the encrypted result as the indicative vector for knapsack set. The NP computation of knapsack problem is applied to make the fragile watermarking system secure.This scheme can effectively resist the birthday attack with a higher resolution than the algorithm before. Theoretical analysis and experimental results demonstrate its effectiveness in resisting the birthday attack and good property of localization.

1 Introduction

Multimedia digitalization provides the accessing of multimedia information with great convenience and also improves the efficiency and accuracy of information expression. With the increasing popularization of Internet, multimedia information communication has stretched to an unprecedented depth and width with more publishing methods. Nowadays, people can issue their own works, important information, and conduct network trade and etc. However, there are many severe problems coming with the popularization of Internet such as easier torts and more convenient tamper. Therefore, how to make full use of the convenience of Internet and also provide effective protection for intellectual property rights has become the focus of research. On this occasion, digital watermarking appears as an effective approach to solve this problem. There are two kinds of digital watermarking, one is robust watermarking and the other is fragile watermarking. Robust watermarking can protect the copyright of digital media and fragile watermarking can verify the integrity and authenticity of digital multimedia. In this paper, we focus on the fragile watermarking.

Recently, many image fragile watermarking techniques have been proposed, such as [1][2][3][4]. Among them, Wong [3][4] proposed to partition an image into sub-blocks and use cryptographic hash function, such as MD5 to compute the digital signature of each image block and the size of the whole image. With the signature as authentication information the scheme embeds watermark into the LSB of every image block. However, the size of image blocks is restricted by

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 771–776, 2004.

the length of the digital signature generated by cryptographic hash function and even if the size of image block is 8×8, it still can not resist the famous birthday attack. In [5] it proposed to enlarge the size of image blocks to resist this attack, and in [6] it proposed to compute the digital signature of an image block and its neighbor blocks to resist this attack. However, these methods both lead to the loss of local resolution.

In this paper, we propose a new method to solve this problem. We apply the intractability of knapsack problem and generate a knapsack set. We encrypt the bit planes of image blocks with random sequence and use the encrypted sequence as the indicative vector for knapsack set to calculate the sum of knapsack sequence. Then we use the result as authentication information. The rest of this paper is structured as follows. In Section 2 we describe the embedding algorithm. The process of extraction and verification will be presented in Section 3. In Section 4 the security of the scheme will be discussed. In Section 5 experimental results are shown and lastly in Section 6 conclusions are given.

2 Embedding Algorithm

First of all, since the knapsack problem will be used to design the fragile watermarking, it is necessary to give a brief introduction on the knapsack problem. The knapsack problem is as follows: given a set $\{a_1, a_2, \cdots, a_n\}$ of positive integers and a sum $s = \sum\limits_{i=1}^{n} x_i a_i$, where each $x_i \in \{0, 1\}$, recover the x_i. It is well known that this problem is NP-complete. For large knapsack problem it is hard to solve. In [7] it is showed that the knapsack problem is NP-complete. At present time the method that can break the knapsack system in polynomial time does not exist.

We describe in the section our fragile watermarking algorithm for grayscale images. For color images the same technique can be applied independently to the color planes of the image, either in the RGB color space or in any other color space such as YUV. Let A denote the original $m \times n$ grayscale image. Then, image A is partitioned into non-overlapping $k \times l$ sub-blocks. These sub-blocks are arranged in raster-scan sequence with A_i denoting the i^{th} block($i = 1, 2, \cdots \left\lceil \frac{m \times n}{k \times l} \right\rceil$). We use $K1$ as the private key to generate a one-dimensional binary vector of length $7 \times m \times n$ (its elements are 0 or 1) and dividing it into one-dimensional sub-vectors r_i of length $7 \times k \times l$, corresponding to the image sub-block A_i. They will be applied to encrypt the bit planes of image blocks. And then we generate a knapsack set $bag = \{a_1, a_2, \cdots, a_{7 \times k \times l}\}$, (Let $\sum\limits_{i=1}^{7 \times k \times l} a_i$ satisfy $\sum\limits_{i=1}^{7 \times k \times l} a_i \leq 2^{k \times l}$, so that the authentication information generated by it can be inserted into the LSB of image blocks.) and set it as public information for verification just like a cryptographic public key.

The following paragraphs will describe the embedding method for one image block.

Step1. According to row priority order, the 7 MSB of image block A_i is arranged into a one-dimensional binary vector of size $7 \times k \times l$. Let v_i denote it.

Step2. We combine r_i with v_i using an exclusive or function. That is, we compute $p_i = v_i \oplus r_i$ where $\oplus$ denotes the element-wise exclusive OR operation between v_i and r_i.

Step3. We use p_i as the indicative vector for knapsack set and calculate the sum $sum_i = \sum_{j=1}^{7\times k\times l} p_i(j) \cdot a_j$. The sum_i is used as authentication information. And then we use $k \times l$ bits to represent sum_i and arrange them into a $k \times l$ binary matrix S_i according to row priority order.

Step4. We combine S_i with W_i using an exclusive or function. That is, we compute $\bar{W}_i = S_i \oplus W_i$ where $\oplus$denotes the element-wise exclusive OR operation between the two vectors. We encrypt $\bar{W}_i$ with the private key K, $C_i = E_k(\bar{W}_i)$. Inserting C_i into LSB of A_i bit by bit.

3 Watermark Extraction and Verification

The sum_i is calculated by the encrypted bit planes of image blocks and knapsack set. The knapsack set, the public information for verification, doesn't change. And the change of pixel grayscale level will absolutely lead to the change of image block's bit planes. Therefore, when pixel grayscale level is changed sum_i will also change. Thus, we can find out whether pixel grayscale level changes or not by only detecting whether sum_i, that is S_i, is changed or not. We partition the image into blocks just like embedding algorithm and split the verification image block V_i into two parts G_i and Q_i. G_i is the LSB part of image block and Q_i is the 7 MSB of the image block. After the same processing in embedding algorithm exerted on Q_i, we first obtain the sum sum'_i and then transfer sum'_i into a binary matrix S'_i. We decipher G_i with private key K, $U_i = D_k(G_i)$ and compute $O_i = S'_i \oplus U_i$ using an element-wise exclusive or procedure. If O_i equals the original watermark block W_i, it means the sum of knapsack sequence is not changed and the image is not tampered either. Conversely, the image is tampered.

4 Security Analysis

Since the same procedure is used to embed the watermark into every image block, we can choose any image block for security analysis and take the i^{th} block for example. Its authentication information is sum_i, $sum_i = \sum_{j=1}^{7\times k\times l} p_i(j) \cdot a_j$, a_j is the element of knapsack set, $p_i = v_i \oplus r_i$. Supposing that an attacker knows sum_i, if the attacker wants to tamper image block A_i successfully, he, first of all, has to build another indicative vector p'_i to satisfy $sum_i = \sum_{j=1}^{7\times k\times l} p'_i(j) \cdot a_j$. However, if the attacker wants to build p'_i, it means that he must solve a knapsack problem containing $7 \times k \times l$ elements. The complexity of this computation is

$2^{7\times k\times l}$ and moreover, p'_i does not necessarily exist. Even if he can build the indicative vector p'_i, the chance of successful tampering is still very small. Because in our scheme the indicative vector is calculated by encrypting the 7 MSB of the image block with random sequence r_i, when the attacker tampers image block A_i into A'_i, that is, v_i is changed into v'_i (The v_i and v'_i are generated by the 7 MSB of A_i and A'_i respectively), he has to satisfy $v'_i \oplus r_i = p'_i$. But he does not know the sub-vector r_i, the random sub-sequence generated by private key $K1$. Because the sub-vector r_i is $7 \times k \times l$ long, there are the $2^{7\times k\times l} 7 \times k \times l$-tuples of 0's and 1's. Therefore, even if the attacker can calculate p'_i, the probability of successful attack will reduced to $1/2^{7\times k\times l}$. Obviously the probability of successful attack will decrease rapidly as $k \times l$ becomes larger. Furthermore, attackers don't know the sum_i, because we encrypt the combination of binary matrix S_i and watermark sequence with the private key K.

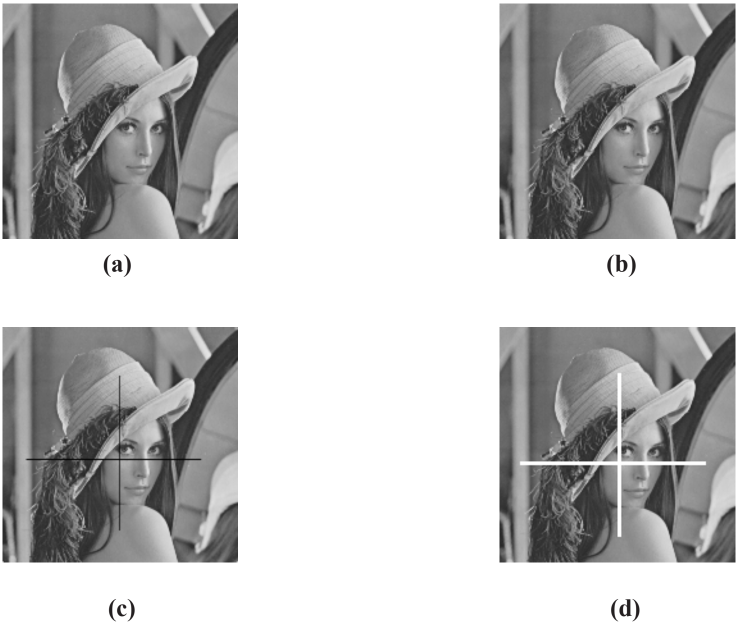
5 Experimental Results

To demonstrate the feasibility of our scheme and high local resolution, two tests have been done. One is locating the tamper area with Wong's scheme proposed in [3,4]. The other is locating the same tamper area with our scheme. The famous Lena image of size 512 × 512 is used as the test image.

In the first test, we use the method of Wong's proposed in [3,4]. The experimental results are shown in Fig. 1(a)-(d). Fig. 1(a) is the original 512 × 512 gray-level Lena image. The watermarked image is given in Fig. 1(b). The tampered watermarked image, where we draw a cross on the face of the girl is shown in Fig. 1(c). Fig. 1(d) is the detection result of tampered area. It indicates that this scheme can successfully detect the tamper. (The white area in the detection result means the places where the image has been tampered.)

In the second test, we use our scheme. With the method we described above, we partition the image into image blocks of size 6×6 equally and randomly select $7 \times 6 \times 6 = 252$ prime numbers that are less than 100,000,000 as the knapsack set. Private key symmetric encryption function has been applied to encrypt the bits series, which generated by the combination of the watermark and the authentication information. (Certainly, we can employ some more complicated encryption algorithms to encrypt the bits series to make the bits series more secure.) Experimental results are shown in Fig. 2(a)-(d). Fig. 2(a) is the original 512×512gray-level Lena image. The watermarked image is given in Fig.2 (b). We can see that the watermarked image has good quality with a peak-signal-noise rate (PSNR) of 51.0db.The tampered watermarked image, where we draw a cross on the face of the girl is shown in Fig. 2(c). Fig. 2 (d) is the detection result of tampered area. It indicates that our scheme can successfully detect the tamper.

From the results above, we can get the following conclusions. Both schemes have a good visual effect, but the local resolution of our scheme is higher than Wong's scheme. (Obviously, the white area in the detection result of our scheme is smaller than that of Wong's.)

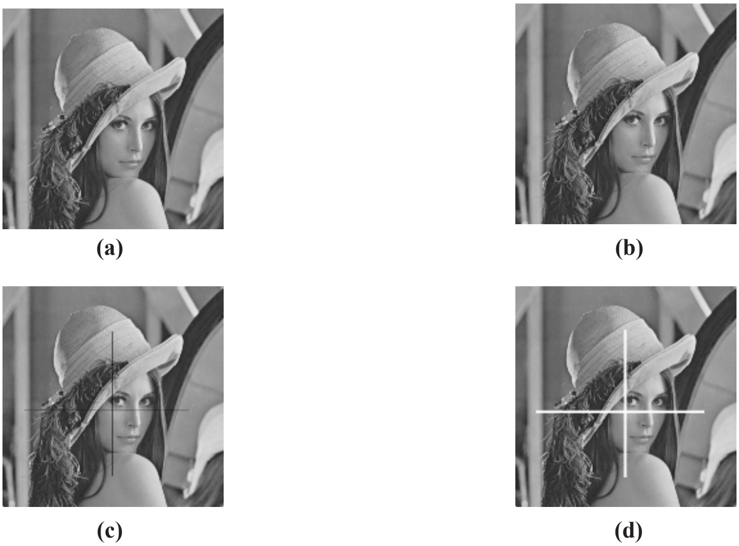

(a)

(b)

(c)

(d)

Fig. 1. The results of Wong's scheme (a) Original image (b) Watermarked image (c) Tampered watermarked image (d) Detection result


(a)

(b)

(c)

(d)

Fig. 2. The results of our scheme (a) Original image (b)Watermarked image (c) Tampered watermarked image (d) Detection result

6 Conclusion

In this paper, we propose a new fragile watermarking for image authentication. The image authentication method is based on the classic NPC problem knapsack problem in algorithm. The scheme can verify every image block by detecting the variation of indicative vector for knapsack set, thus the tampered area can be located. It applies the NP computation of knapsack problem to resolve the security problem existing in the common methods. Analysis and experimental results demonstrate its effectiveness and practicability.

References

1. M. Yeung and FC Mintzer: An Invisible Watermarking Technique for Image Verification. International Conference on Image Processing, Vol.2. Santa Barbara, USA. (1997) 680-683
2. R. B. Wolfgang, E. J. Delp: Fragile Watermarking Using the VW2D Watermark. The SPIE/IS&T International Conference on Security and Watermarking of Multimedia Contents. Vol. 3657. San Jose, California (1999) 204-213
3. P. W. Wong: A Public Key Watermark for Image Verification and Authentication. IEEE International Conference on Image Processing, Vol.1. Chicago, Illinois, USA (1998) 455-459
4. P. W. Wong: A Watermark for Image Integrity and Ownership Verification. IS&T PIC Conference. Portland, Oregon, USA (1998) 374-379
5. Paulo SLM Barreto and Hae Yong Kim: Pitfalls in Public Key Watermarking. XII Brazilian Symposium on Computer Graphics and Image Processing. Campinas, SP-Brazil (1999) 241-242
6. Paulo S. L. M. Barreto, Hae Yong Kim, Vincent Rijmen: Toward A Secure Public-key Blockwise Fragile Authentication Watermarking. IEEE International Conference on Image Processing. Thessaloniki, Greece (2001) 494-497
7. M. R. Garey, D. S. Johnson: Computers and Intractability: A Guide To The Theory of NP-completeness. WH Freeman and Company, San Francisco (1979)

Author Index