Lecture Notes in Computer Science 3893

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Moshe Y. Vardi
Rice University, Houston, TX, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Luigi Atzori   Daniel D. Giusto
Riccardo Leonardi   Fernando Pereira (Eds.)

Visual Content Processing and Representation

9th International Workshop, VLBV 2005
Sardinia, Italy, September 15-16, 2005
Revised Selected Papers

Springer

Volume Editors

Luigi Atzori
Daniel D. Giusto
University of Cagliari, DIEE
Piazza d'Armi, 09123 Cagliari, Italy
E-mail: l.atzori@diee.unica.it,ddgiusto@unica.it

Riccardo Leonardi
University of Brescia, DEA
Via Branze 38, 25123 Brescia, Italy
E-mail: riccardo.leonardi@ing.unibs.it

Fernando Pereira
Instituto Superior Técnico, Instituto de Telecomunicações
Torre Norte, Sala 10.14, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
E-mail: fernando.pereira@lx.it.pt

Library of Congress Control Number: 2006923376

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

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

ISSN 0302-9743
ISBN-10 3-540-33578-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33578-8 Springer Berlin Heidelberg New York

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

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11738695 06/3142 5 4 3 2 1 0

Preface

This book is a post-conference publication and contains a selection of papers presented at the VLBV workshop in 2005, which was held in Sardinia, Italy, on September 15-16.

VLBV 2005 attracted very high interest among the scientific community, and the workshop received 85 submissions. Fifty-five papers were accepted and presented in three poster sessions focusing on "New Perspectives in Handling Multimedia Data", "Wavelet Based Compression: Advances and Perspectives", and "Applications of Distributed Video Coding".

Beyond providing a forum for the presentation of high-quality research papers in various complementary aspects of visual content processing and distribution, the workshop gave the opportunity of exchanging ideas and opinions on hot topics in the fields by means of three open panels. The first was devoted to the role of standards, with the attempt of both panelists and audience to provide answers to challenging questions on the success of standards, their industrial spinoff, and the involvement of academic people in standardization activities. The second panel aimed to discuss the advances and the perspectives of wavelet based compression. The third one focused on the applications of distributed video coding; the discussion mainly regarded the real performance of such an approach in video coding when compared with traditional approaches.

In addition to these components, the program was enriched by two stimulating invited talks given by world-renowned researchers in the field: "Digital Cinema", by Michael W. Marcellin and "Distributed Video Coding", by Bernd Girod.

We would like to thank all the authors for submitting their manuscripts to this workshop and express our special thanks to all the steering committee members and the external reviewers for their work in reviewing the papers.

Finally, we would like to express our gratitude to all the members of the organizing committee for their dedicated work and to the representatives of our sponsoring organizations for their continuous support.

For more information on VLBV 2005 visit: www.diee.unica.it/vlbv05

December 2005

Luigi Atzori
Daniele D. Giusto
Riccardo Leonardi
Fernando Pereira

Organization

VLBV 2005 was organized by the Department of Electrical and Electronic Engineering, University of Cagliari, Italy.

Executive Committee

General Chairs:	Daniel D. Giusto (University of Cagliari, Italy) Riccardo Leonardi (University of Brescia, Italy)
Technical Chairs:	Luigi Atzori (University of Cagliari, Italy) Fernando Pereira (Instituto Superior Técnico, Portugal)

Steering Committee

Kiyoharu Aizawa, University of Tokyo, Japan
Jan Bormans, IMEC, Belgium
Leonardo Chiariglione, Digital Media Project, Italy
Reha Civanlar, Koç University, Turkey
Touradj Ebrahimi, EPFL, Switzerland
Narciso Garcia, Grupo de Tratamiento de Imágenes, Spain
Bernd Girod, Stanford University, USA
Thomas Huang, University of Illinois, USA
Aggelos Katsaggelos, Northwestern University, USA
Stefanos Kollias, National Technical University of Athens, Greece
Philippe Salembier, Universitat Politècnica de Catalunya, Spain
Ralf Schaefer, Heinrich-Hertz Institut, Germany
Thomas Sikora, Technische Universität Berlin, Germany
Gary J. Sullivan, Microsoft, USA
Murat Tekalp, Koç University, Turkey, and University of Rochester, USA
Antony Vetro, MERL, USA
Avideh Zakhor, University of California at Berkeley, USA

Sponsoring Institutions

Tiscali, Italy
Akhela, Gruppo Saras, Italy
DIEE, University of Cagliari, Italy

Table of Contents

Expected Distortion of Dct-Coefficients in Video Streaming over Unreliable Channel

Marco Fumagalli[1], Marco Tagliasacchi[2], and Stefano Tubaro[2]

[1] CEFRIEL - Politecnico di Milano, Via R. Fucini 2 - 20133 Milano, Italy
fumagall@cefriel.it
[2] Dip. di Elet.e Inf., Politecnico di Milano, P.zza L. Da Vinci, 32 - 20133 Milano, Italy
{tagliasacchi, tubaro}@elet.polimi.it

Abstract. The Recursive Optimal per-Pixel Estimate (ROPE) algorithm allows the encoder to estimate the pixel-by-pixel expected distortion of the decoded video sequence due to channel loss. The algorithm requires in input an estimate of the packet loss rate and the knowledge of the error concealment technique used at the decoder with no need to perform any comparison between original and decoded frames. Although the ROPE algorithm computes the expected distortion in the pixel domain, in some applications it is important to have access to the expected distortion in the DCT domain, e.g., for an accurate allocation of the redundancy bits in error-resiliency schemes. This paper presents the extension of the ROPE algorithm in the transform DCT domain that allows estimating the expected distortion of the decoded video sequence for each DCT coefficient.

1 Introduction

Nowadays, sending a video sequence over a network that does not provide any QoS guarantee (e.g., IP network) is a very common application. If errors occur, some information does not reach the decoder and error-concealment techniques cannot completely avoid error propagation due to inter-frame dependency introduced by predictive encoding. In the design of error-resilient approaches, it is interesting for the sender to estimate the decoded video distortion that the receiver is expected to suffer.

The Recursive Optimal per-Pixel Estimate (ROPE) algorithm [2] allows the encoder to calculate the pixel-by-pixel expected distortion of the decoded video due to channel loss. In the case of packet-switched networks, the only required parameters are an estimate of the network Packet Loss Rate (PLR) and the error concealment technique used at the decoder side. The encoder does not know the loss pattern, hence it has to characterize the actual reconstruction of a pixel value operated by the decoder as a random variable. The algorithm in [2] is based on the assumption that the considered video sequence is encoded with integer-pixel motion vectors. Since most of the common encoders implement a sub-pixel precision motion estimation, the work in [1] proposes an extension of the ROPE algorithm in order to obtain an accurate solution for this case.

The knowledge of the expected distortion in the reconstructed video sequence is a valuable information in several applications e.g., when comparing different encoding

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 1 – 8, 2006.
© Springer-Verlag Berlin Heidelberg 2006

techniques in various scenarios (in [3] the competitors are MDC and optimized one-layer encoding), tuning of the encoder parameters (in [2, 4, 6] an estimate of the decoded video distortion, due to packet losses, is used for mode decision and to improve the error resilience of the stream), etc. In all these approaches the expected distortion is computed in the pixel domain and applied at macro-block level.

In contrast with the above-mentioned approaches, in several emerging applications, it can be interesting to have access to an estimate of the expected distortion directly in the DCT domain. This information can be used, e.g., for an accurate allocation of the redundancy in error-resilience schemes. More recently, error resiliency tools based on the principles of distributed source coding have appeared in the literature [5, 7, 8]. In [8] an estimate of the expected distortion for each DCT coefficient is used to drive the rate allocation of the side channel, i.e. a redundant representation that is used to correct errors at the decoder, thus stopping drift. The fundamental idea behind the work in [8] is that the encoder needs to characterize in statistical terms the induced channel noise between the original sequence and the reconstructed sequence at the decoder (the so-called side-information in distributed source coding jargon). Intuitively, the higher is the estimated noise level, the larger is the number of bits that need to be spent for the side channel in order to correct errors. In [8], a simple approximation of the DCT coefficients distortion is given by an extension of ROPE algorithm to the transform domain. Simulation results reveal that this approach leads to a coarse approximation of the estimated distortion. In this work we propose a more accurate algorithm to estimate the expected decoded distortion of each DCT coefficient.

This paper is organized as follows: Section 2 presents a brief overview of ROPE algorithm and its extension to half-pixel precision motion estimation. In Section 3 the DCT extension of ROPE presented in [8] is briefly summarized. Section 4 describes the proposed video distortion estimation algorithm (EDDD) in the DCT domain. Section 5 shows some simulation results and the conclusions are given in Section 6.

2 Rope Algorithm Overview

In this section the original ROPE algorithm [2] is briefly summarized together with its extension to half-pixel precision motion-estimation as proposed in [1].

Integer-pixel ROPE

The expected distortion at the decoder for pixel i in frame n can be expressed as in Eq. (1)

$$d_n^i = E\left\{\left(f_n^i - \tilde{f}_n^i\right)^2\right\} = \left(f_n^i\right)^2 - 2f_n^i E\left\{\tilde{f}_n^i\right\} + E\left\{\left(\tilde{f}_n^i\right)^2\right\} \tag{1}$$

where f_n^i is the value of pixel in original video, and $\tilde{f}_n^i$ is the decoder reconstruction; p is the probability of a packet to get lost. We assume this quantity is known at the encoder side. According to Eq. (1), $\tilde{f}_n^i$ first and second moments are required to compute the distortion. We assume a simple motion-compensated temporal error-concealment scheme, i.e. the lost MB is replaced with the one in the previous frame pointed by the motion vector (MV) of the above MB; if not available, a simple zero-motion replacement is used.

For an Inter-coded MB (similar equations can be written for the Intra case [2]), the ROPE algorithm computes recursively the first and second moments, frame after frame, for each pixel, as in Eq (2) and Eq. (3).

$$E\left\{\tilde{f}_n^i\right\} = (1-p)\left(\hat{e}_n^i + E\left\{\tilde{f}_{n-1}^j\right\}\right) + p(1-p)E\left\{\tilde{f}_{n-1}^k\right\} + p^2 E\left\{\tilde{f}_{n-1}^i\right\} \tag{2}$$

$$\begin{aligned} E\left\{\left(\tilde{f}_n^i\right)^2\right\} = (1-p)\left(\left(\hat{e}_n^i\right)^2 + 2\hat{e}_n^i E\left\{\tilde{f}_{n-1}^j\right\} + E\left\{\left(\tilde{f}_{n-1}^j\right)^2\right\}\right) + \\ + p(1-p)E\left\{\left(\tilde{f}_{n-1}^k\right)^2\right\} + p^2 E\left\{\left(\tilde{f}_{n-1}^i\right)^2\right\} \end{aligned} \tag{3}$$

$\hat{f}_n^i$ is the correct reconstruction of the considered pixel and $\hat{e}_n^i$ is the reconstruction of the prediction error (*i* index refers to actual pixel location, *j* to the location pointed by the actual MV and *k* by the reconstructed MV).

If the video is encoded at half-pixel precision, the algorithm proposed in [2] rounds the half-pixel precision MVs to integer values. This solution leads to a sub-optimal distortion estimate that is often unacceptable.

Half-pixel ROPE

In [1] the ROPE algorithm is extended to deal with half-pixel precision MVs. The expression of a half-pixel *j*, interpolated horizontally or vertically from two adjacent integer-pixels j_1 e j_2, is reported in Eq. (4)

$$f_{n-1}^j = \left\lfloor \frac{r + f_{n-1}^{j_1} + f_{n-1}^{j_2}}{2} \right\rfloor \tag{4}$$

where $\lfloor \ \rfloor$ represents the integer part and *r* is a polarization terms (to obtain a zero-mean quantization error). Eq. (5) and (6) present the expressions of the first and second moments of Eq. (4), neglecting the integer rounding.

$$E\left\{\tilde{f}_{n-1}^j\right\} = \frac{r + E\left\{\tilde{f}_{n-1}^{j_1}\right\} + E\left\{\tilde{f}_{n-1}^{j_2}\right\}}{2} \tag{5}$$

$$\begin{aligned} E\left\{\left(\tilde{f}_{n-1}^j\right)^2\right\} = \frac{1}{4}\Big(r^2 + E\left\{\left(\tilde{f}_{n-1}^{j_1}\right)^2\right\} + E\left\{\left(\tilde{f}_{n-1}^{j_2}\right)^2\right\} + \\ + 2r\left(E\left\{\tilde{f}_{n-1}^{j_1}\right\} + E\left\{\tilde{f}_{n-1}^{j_2}\right\}\right) + 2E\left\{\tilde{f}_{n-1}^{j_1}\tilde{f}_{n-1}^{j_2}\right\}\Big). \end{aligned} \tag{6}$$

In Eq. (5) and (6) every term is known but the last one in Eq. (6), which represents the expectation of the product of two adjacent integer pixels[1]. If the two pixels j_1 and j_2 of frame *n* belong to the same MB, in every possible packet loss pattern they have the same MV. The work in [1] proposes to calculate the expected value of the product

[1] The work in [6] proposes to approximate the expectation of the product of two adjacent integer pixels with its superior limit indicated by Cauchy-Schwartz inequality.

of two integer adjacent pixels as a linear combination of the various expected values of products of the adjacent pixels from whom they are predicted in the different loss scenarios as in Eq. (7) (analogue equations for the Intra case can be found in [1]).

$$E\left\{\tilde{f}_n^{i_1}\tilde{f}_n^{i_2}\right\}=(1-p)\cdot \tag{7}$$
$$\left(\hat{e}_n^{i_1}\hat{e}_n^{i_2}+\hat{e}_n^{i_1}E\left\{\tilde{f}_{n-1}^{j_2}\right\}+\hat{e}_n^{i_2}E\left\{\tilde{f}_{n-1}^{j_1}\right\}+E\left\{\tilde{f}_{n-1}^{j_1}\tilde{f}_{n-1}^{j_2}\right\}\right)+$$
$$+p(1-p)E\left\{\tilde{f}_{n-1}^{k_1}\tilde{f}_{n-1}^{k_2}\right\}+p^2E\left\{\tilde{f}_{n-1}^{i_1}\tilde{f}_{n-1}^{i_2}\right\}.$$

If the MV associated to the pair of integer pixels i_1 and i_2 is a half-pixel precision MV, the reference pair of pixel is formed by half-pixels (called j_1 and j_2). A possible configuration is shown in Figure 1.

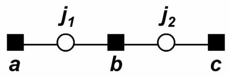


Fig. 1. Two horizontally adjacent pixels (j_1 and j_2) that are used as references. Their values come from interpolation of the consecutive pixels a, b, and c.

In order to calculate the expected value of the product of half-pixels j_1 and j_2, given the known expected products between integer (a,b) and (b,c), the product can be developed substituting j_1 and j_2 with their expression as in Eq. (4). In this way a new unknown term is needed: the expected value of the product between not adjacent integer pixels a and c. The problem is tackled using the well-known relation of Eq. (8), where μ e σ represent the expected value and the standard deviation of reconstructed pixel - known by Eq. (2) and (3) - and $\rho_{a,c}$ is the correlation coefficient between $\tilde{f}_{n-1}^{a}$ e $\tilde{f}_{n-1}^{c}$; $\rho_{a,c}$ is then approximated (assuming an autoregressive linear model between adjacent pixels) by the expression in Eq. (9).

$$E\left\{\tilde{f}_{n-1}^{a}\tilde{f}_{n-1}^{c}\right\}=\rho_{a,c}\sigma_a\sigma_c+\mu_a\mu_c \tag{8}$$

$$\rho_{a,c}=\rho_{a,b}\cdot\rho_{b,c} \tag{9}$$

In case pixels i_1 and i_2 belong to different MBs (it happens less than 10% of the cases), they are supposed to be uncorrelated and the expected values of products of the adjacent pixels are approximated by the product of the expected values. The results presented in [1] show the accuracy of this solution on the whole sequence.

3 DCT-ROPE

In [8] an extension of ROPE that works directly in the DCT domain is proposed. The basic idea is to use the recursive equation of the original ROPE algorithm [2] on the DCT coefficients instead of working on pixel values. This straightforward solution (that we call here DCT-ROPE) has to cope with the difficulty of managing the motion compensation phase in the transform domain.

In order to map DCT coefficients from the current to the reference frame, motion vectors are quantized with a step size equal to the block side length in such a way that each macro-block of the current frame is matched with the nearest macro-block of the reference frame. This coarse approximation of motion vectors leads to a loss of accuracy in the distortion estimation as shown by the simulation results reported in Section 5.

4 Expected Distortion of Decoded Dct-Coefficients (EDDD)

This section illustrates the proposed algorithm that allows estimating the distortion of the reconstructed video sequence in the DCT domain, yet retaining the accuracy of the original ROPE algorithm. The basic idea of is to run ROPE in the pixel domain and then to estimate the distortion in the DCT domain, block-by-block, using the corresponding spatial information. This is rather different from the approach presented in [8], where the estimate was obtained directly in the DCT domain.

As we saw in the previous section, the ROPE algorithm considers the decoded value of each pixel j as a random variable x_j whose statistics are expressed by the estimated first and second moments of the pixel value after the reconstruction at the decoder side. The two expected quantities are calculated by recursive equations. In similar way, for each image block, we represent each of the i-th DCT coefficient as a statistical variable y_i. In order to characterize the expected distortion we need to estimate both the first ($E[y_i]$) and the second ($E[y_i^2]$) moments. In the following, we refer to the proposed approach as Expected Distortion of Decoded DCT-coefficients (EDDD).

Given w_{ji} (the j-th elements of the i-th DCT basis function) the random variable y_i can be written as in Eq. (10).

$$y_i = \sum_J w_j^i \cdot x_j \tag{10}$$

Using Eq. (10) we obtain the expression of $E[y_i]$ and $E[y_i^2]$ as reported in Eq. (11) and Eq. (12).

$$E[y_i] = \sum_J w_j^i \cdot E[x_j] \tag{11}$$

$$E[y_i^2] = \sum_Z \sum_J w_z^i \cdot w_j^i \cdot \left(E[x_z x_j] - E[x_z] \cdot E[x_j]\right) + E[y_i]^2 \tag{12}$$

Eq. (11) is of easy calculation while Eq. (12) presents an unknown term that represents the expected value of the product of pixels pairs within the considered block. Although this quantity is not needed in the full pixel precision version of ROPE [2], the work in [1] introduces an extension of ROPE to work with half-pixel precision, as briefly summarized in Section 2. In particular, the additional recursive Eq. (7) gives the expected value of the product of adjacent pixels in both horizontal and vertical directions. The values of $E[x_z\ x_j]$ for not adjacent pixels z and j are obtained by the combination of the expected values of the product of the adjacent pixels that connect z with j. This latter approximation is the only cause of inaccuracy of the proposed EDDD algorithm.

Besides its accuracy, the proposed approach has the significant feature of energy preservation over each block, i.e., the distortion calculated by ROPE algorithm over each block in pixel domain is exactly the same distortion calculated by EDDD algorithm in the DCT domain. In fact it is easy to show that this feature is not affected by the aforementioned simplification and thus it is always true.

5 Experimental Results

This section compares the proposed EDDD approach with the ROPE algorithm applied directly in the DCT domain (DCT-ROPE). In order to assess the absolute accuracy and consistency of both approaches, they are compared with the actual distortion of the reconstructed sequence averaged over several network simulations with different error patterns.

Figure 2 illustrates the PSNR tracks at frame level of the decoded 'Foreman' sequence (QCIF, 30 fps, 256 kbps, Intra MB refresh 10%) subjected to a random packet loss rate of 10%. The network simulations track is the average PSNR over 60 different realizations, while pixel-domain ROPE (PEL-ROPE) and EDDD have the same track due to the energy conservation feature of EDDD. The inaccuracy of DCT-ROPE approach is evident. Similar results can be obtained over other video sequences.

Figure 3 shows the accuracy of EDDD and DCT-ROPE algorithms in the DCT domain. Each 8x8 DCT block is divided into four not-overlapped frequency bands:

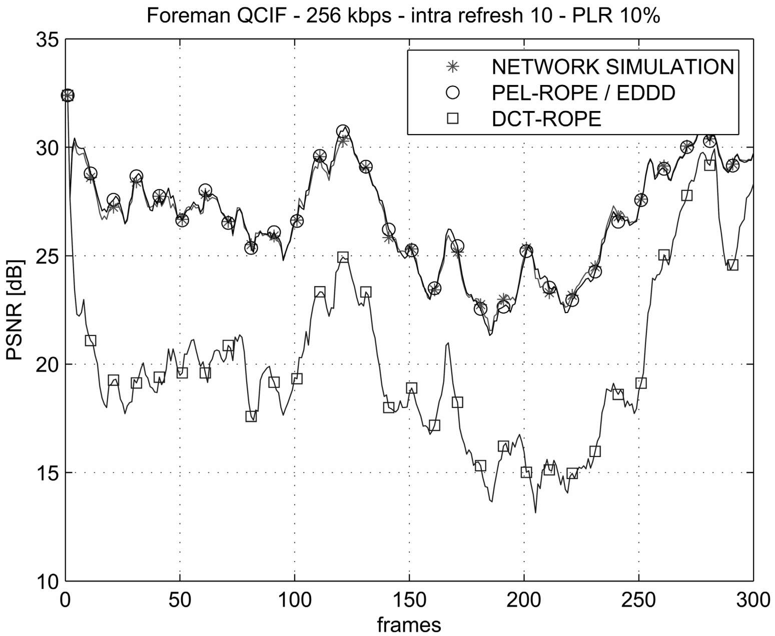


Fig. 2. PSNR tracks for EDDD, DCT-ROPE and NS

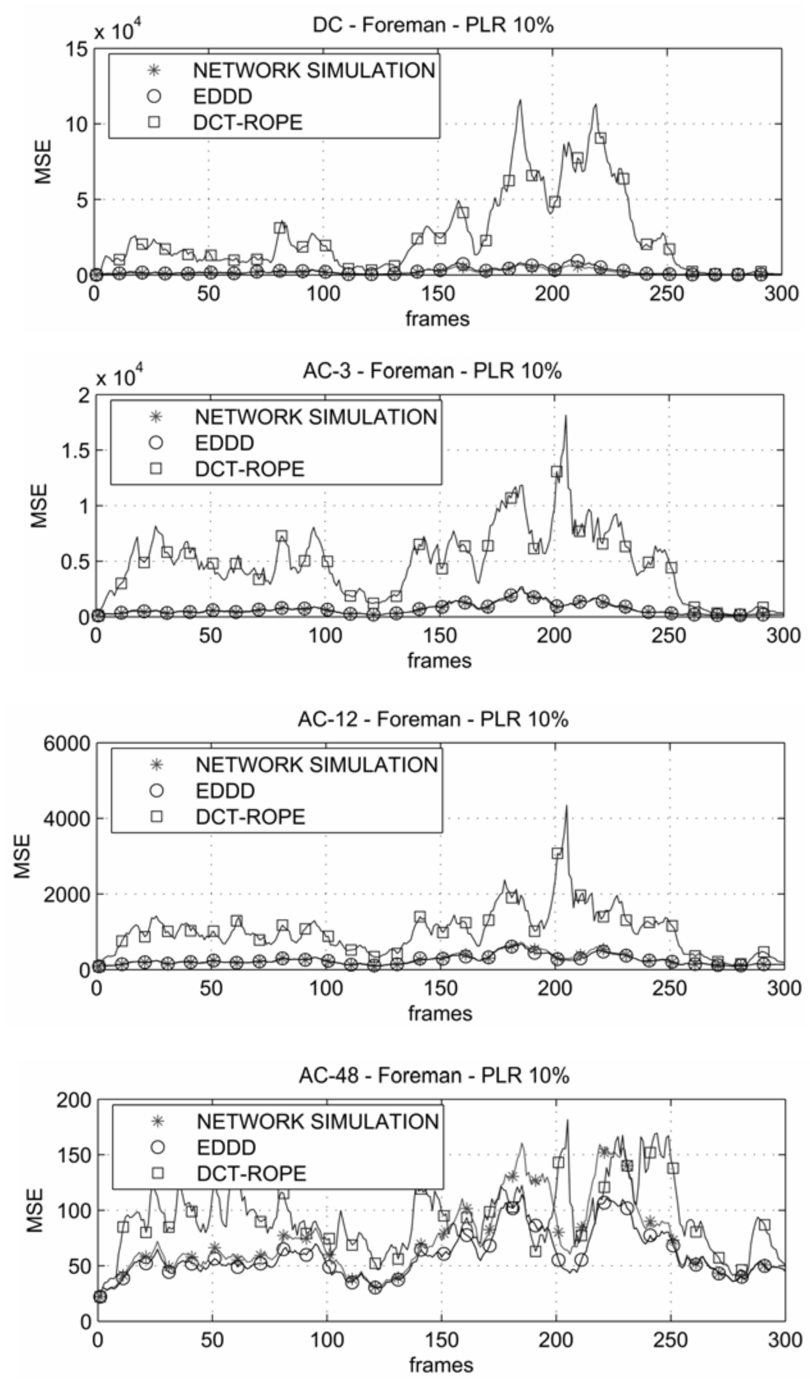


Fig. 3. MSE in the four not-overlapped DCT bands

DC coefficient, AC-3 (the three coefficients adjacent to DC), AC-12 (the twelve coefficients around AC-3), and AC-48 (the remaining coefficients). For these bands we calculate the expected distortion (MSE). As we can see, the proposed EDDD approach presents a significant accuracy improvement with respect to the DCT-ROPE approach. Moreover the estimated distortion is rather close to the actual average distortion measured in the network simulations Tests on other video sequences at different PLRs lead to the same results.

To conclude, we showed that the proposed EDDD approach in DCT domain presents the same expected distortion of ROPE at frame level and that it estimates with considerable accuracy the expected distortion in the DCT domain.

6 Conclusions

This paper presents an extension of ROPE algorithm that is able to estimate the expected distortion in the DCT domain. The basic idea of the proposed EDDD approach is to run the original ROPE algorithm in the pixel domain and then to estimate the DCT domain, block-by-block, using the corresponding spatial information. The accuracy of the EDDD is validated by several simulation results.

References

1. Bocca, V., Fumagalli, M., Lancini, R., Tubaro, S.: Accurate Estimate of the Decoded Video Quality: Extension of ROPE Algorithm to Half-Pixel Precision. Proc. of PCS 2004, San Francisco (2004)
2. Zhang, R., Regunathan, S.L., Rose, K.: Video coding with optimal inter/intra mode switching for packet loss resilience. IEEE JSAC **18** (June 2000)
3. Reibman, A.: Optimizing multiple description video coders in a packet loss environment. Proc. of 12th PV Workshop, Pittsburgh (2002)
4. Zhang, R., Regunathan, S. L., Rose, K.: Switched error concealment and robust coding decisions in scalable video coding. Proc. IEEE ICIP 2003, Barcelona (Spain), (2003)
5. Sehgal, A., Ahuja, N.: Robust Predictive Coding and the Wyner-Ziv Problem. Proc. DCC 2003, Snowbird, Utah (2003)
6. Leontaris, A., Cosman, P.: Video compression for lossy packet networks with mode switching and a dual-frame buffer. IEEE Trans. on Image Proc. **13** (2004)
7. Rane, S., Aaron, A., Girod, B.: Systematic lossy forward error protection for error-resilient digital video broadcasting - A Wyner-Ziv coding approach. Proc. IEEE ICIP 2004, Singapore (2004)
8. Majumdar, A., Wang, J., Ramchandran, K.: Drift Reduction in Predictive Video Transmission using a Distributed Source Coded Side-Channel. ACM Multimedia, (2004)

Content Adaptation Tools in the CAIN Framework

Víctor Valdés and José M. Martínez

Grupo de Tratamiento de Imágenes,
Escuela Politécnica Superior, Universidad Autónoma de Madrid,
Avda. Francisco Tomás y Valiente, 11, Ciudad Universitaria de Cantoblanco,
Ctra. de Colmenar, Km 15, E-28049 Madrid, Spain
{Victor.Valdes, JoseM.Martinez}@uam.es
http://www-gti.ii.uam.es

Abstract. This paper presents the Content Adaptation Tools currently integrated in the CAIN framework, a content adaptation manager targeted to the integration of different metadata-driven content adaptation approaches. The CAIN system is in charge of managing the available content adaptation tools in order to perform the most appropiate adaptation modality taking account of the available content metadata. Current system architecture and adaptation process are presented detailing the decision process about adaptation parameters and adaptation tool selection taken in order to perform content adapatation. The different Content Adaptation Tools types are presented together with details of the currently integrated adaptation tools.

1 Introduction

Content Adaptation is the main objective of a set of technologies that can be grouped under the umbrella of the Universal Multimedia Access (UMA) concept[1] providing the technologies for accessing to rich multimedia content through any client terminal and network and in any environmental condition and user, always targeting to enhance the user's experince[2].

CAIN (Content Adaptation Integrator)[3] is a content adaptation manager that is designed to provide Metadata-driven content adaptation integrating different but complementary content adaptation approaches[4]: transcoding, transmoding, scalable content, summarization, semantic driven adaptation, …

Key technologies within the CAIN include descriptions and content adaptation tools.

The content descriptions are based on MPEG-7 MDS and MPEG-21 DIA BSD, whilst the context descriptions are based on a subset of MPEG-21 DIA Usage Environment Descriptions tools [3]. The content adaptation tools (CATs) are grouped in different categories depending on their functionality. This paper presents an overview of the current CAIN system architecture and afterwards details the status of the different Content Adaptation Tools already working within CAIN.

2 CAIN System

As shown in Fig. 1 CAIN is integrated by the Decision Module -DM- and several CATs, Encoders and Decoders. The Encoders and Decoders have been introduced in

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 9 – 15, 2006.
© Springer-Verlag Berlin Heidelberg 2006

the architecture due to requirements from the aceMedia Project[5] in order to provide generic "transcoding" combinations via an intermediate raw format and to use CAIN as an input/output node accepting/delivering media in raw format.

In response to an external invocation (via the corresponding API) the CAIN system requests and receives content, MPEG-7 compliant content descriptions and MPEG-21 compliant bit streams descriptions. In parallel, the system receives a context description which contains, at least, media related user preferences, terminal capabilities and network characteristics. All those inputs are parsed and the resultant information is received by the Decision Module which is in charge of confronting the selection of media parameters that would fulfil the adaptation and to decide which of the available CATs, Encoders or Decoders should be used. The chosen CAT/Codec is then launched and managed in order to produce adapted content and metadata which will be transferred to the delivery services.

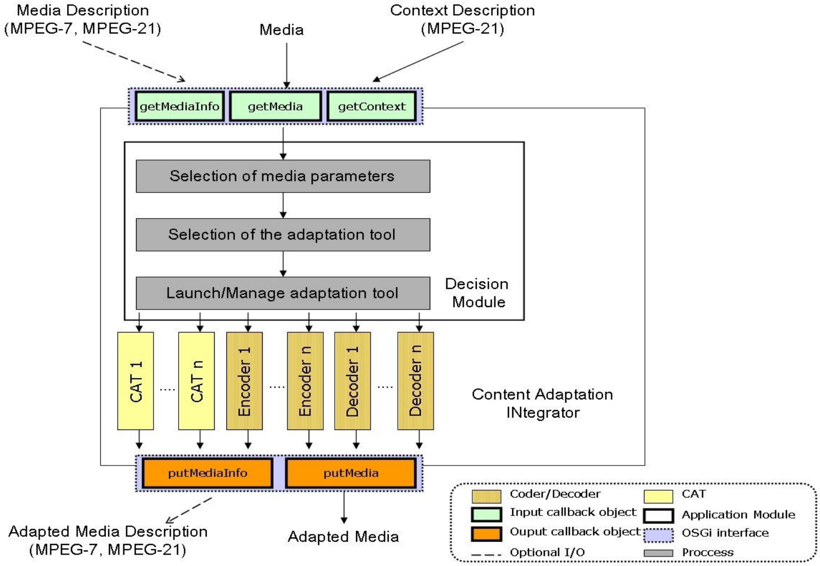


Fig. 1. CAIN Architecture

The CAIN architecture is designed in order to allow the addition of new codecs and CATs. In both cases it's necessary to provide the new CAT/Codec following the CAIN API specifications and a CAT Capabilities description file including information about the input and output formats accepted by the new CAT/Codec and its adaptation capabilities. The Decision Module in charge of selecting the appropriate CAT/Codec to perform the adaptation will check the available CAT Capabilities in order to choose the CAT/Codec able to perform the desired adaptation or the one able to perform the nearest possible adaptation in case of not being able to perform the initially desired one.

Therefore, each new CAT/Codec to be added to CAIN should implement a previously defined API to communicate with the Decision Module and should provide a CAT Capabilities description file. The CAT Capabilities description file (based on

MPEG-7 MDS and MPEG-21 DIA description tools) defines the list of adaptation capabilities specifying in each case which kind of adaptation(s) the CAT is able to perform and the possible list of parameters that define the mentioned adaptation such as input format, output format, and different features depending on which kind of adaptation is being defined: e.g., accepted input/output frame rate, resolution, channels, bitrate, etc... The CAT Capabilities description file should provide all the necessary CAT characteristics to allow the CAIN system to take a decision about which of the available CATs is the best choice to perform an adaptation.

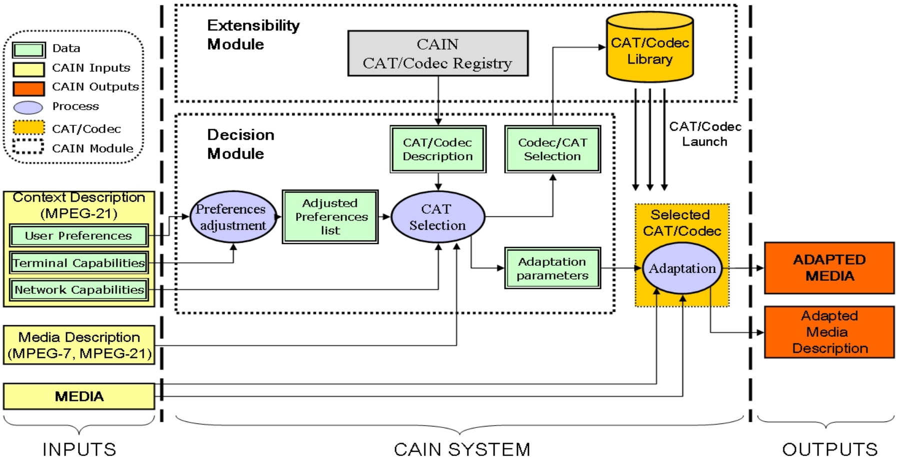


Fig. 2. CAIN Adaptation Process Overview

The CAT Capabilities description file is parsed in order to sign up the CAT/Codec in the CAIN registry, which is necessary for the Decision Module to know that a new CAT/Codec is available and which are its characteristics.

Fig. 2 shows an overview of the adaptation process. In the first step the received and parsed user preferences and terminal capabilities are contrasted in order to get a list of user preferences constrained by the terminal capabilities, the 'Adjusted Preferences List'. In the next step the 'Adjusted Preferences List' the Media Description (MPEG-7/MPEG-21 description) and the network capabilities are used to take a decision about which is the best available CAT/Codec to perform the adaptation. The CAT/Codecs Capabilities descriptions are checked in order to find the list of adaptation tools that can perform the desired adaptation or, in case of not finding a tool which fulfils all the adaptation constraints, a tool which could perform a similar adaptation. It's necessary to check, for each of the adaptation preferences if the adaptation tool is able to adapt the input media fulfilling this condition. Finally the adaptation tool wich is able to perform the most similar adaptation to the 'Adjusted Preferences List' is selected.

Currently, the similarity criterion is based on the most simple metric: the euclidean distance between the adapted media characteristics (such as resolution, frame rate, ...) and the characteristics of the target media format, that is the distance between the ‘User Preferences’ and the ‘Adjusted Preferences List’ in the Preferences Adjustment step (see Fig. 2) and also the distance between the Adjusted Preferences List and the final Adaptation Parameters in the CAT Selection step. The distance between each one of those characteristics is computed as the percentage with respect to original value of the media and afterwards all the individual figures are added to obtain the adaptation similarity figure.

We are studying the use of other metrics and the inclusion of weights for each characteristic. The weights will allow to take into account user preferences (or default ones if those are not provided based on subjective studies to be performed) based on the ConversionPreferences and PresentationPriorityPreferences description tools specified as part of the DIA Usage Environment description tools.

Another factor to be considered is the adaptation cost. In the future implementations of the system each adaptation characteristic will be measured in terms of computational complexity. The final objetive is to get a balance between adaptation fidelity (or similarity) to the user preferences and computational cost of the adaptation process. This system will allow, in case of having several CATs available to perform an adaptation, to choose the most efficient CAT to adapt the content.

As result of this step a CAT/Codec is selected and the definitive set of adaptation parameters are obtained. After this step it's possible that the adaptation parameters obtained would not fit exactly the original user preferences as the adaptation is constrained by the CAT/Codec capabilities. After an adaptation tool has been selected it is invoked to perform the adaptation, receiving the original media, the media description and the definitive adaptation parameters and providing as output the adapted media and the adapted media description.

3 Current CAIN CATs

The CAIN system approximation to UMA is based on the integration of several adaptation tools in order to be able to perform different types of adaptation processes. This approximation allows the system to choose different adaptation tools in each situation aiming to get the most efficient and accurate content adaptation. In this section several of the current available CATs in the CAIN system are described.

Currently there are 4 different possible CAT categories, plus the 2 codec ones:

- ‘Transcoder CATs’ are in charge of classical transcoding.
- ‘Scalable Content CATs’ are in charge of truncation (limited expansion –e.g. interpolation- will be studied for further versions) of scalable content. It uses a format agnostic approach, accessing the bitstreams via a generic bitstream transcoding module (MPEG-21 DIA BSD/gBSD).
- ‘Semantic driven real-time CATs’ are in charge of the extraction, via the use of real time analysis techniques, of semantic features from content in order to generate a semantic-driven adapted version, either at signal level content adaptation (e.g., ROIs) or at content summarization level.
- ‘Transmoding CATs’ are intended to provide different kinds of transmoding (e.g., “simple” audiovisual to audio, video to slide-show transmoding, media to text (or voice) transmoding)
- Encoders are in charge of encoding raw data to a specific coding format.
- Decoders are in charge of decoding a specific coding format to raw data.

3.1 Transcoding CATs

Currently there are two transcoding CATs able to deal respectively with audiovisual content (audio and video separately or jointly) and images

The Audiovisual Transcoding CAT supports transcoding in the pixel domain, setting frame rate, frame size, sampling and other coding parameters. Currently, the Audiovisual Transcoding CAT is able to deal with MPEG-1/2/4 SP video and MPEG-1 layer 2, MPEG-1 layer 3, MPEG-4 AAC and AMR narrowband audio. It's also possible to change between several video/audio container formats such as avi, mpg, mp4, etc, file formats.This CAT is based on the use of ffmpeg[6].

The Image Transcoding CAT supports image format transcoding and resolution adaptation and is able to read/write JPEG, BMP, GIF, TIFF, PPM image formats allowing resolution changes in the image before saving them. JPEG2000 codification is also available from the supported formats but no JPEG2000 to other formats transcoding is supported: once an image is transcoded to JPEG2000 format the Scalable Image CAT will be in charge of this image adaptation in order to take adavantage of the scalability capabilities of this format. This CAT core is based in the Java Advanced Imaging (JAI) API.

3.2 Scalable Content CATs

The usage of MPEG-21 DIA's BSD tool allows format-agnostic adaptation of scalable content. BSD provides mechanisms to perform scalabe content adaptation via media file truncation without prior knowledge of the media file structure. This is based on the availability of xsd and xsl descriptions of the media format and media transformations respectively. Currently, the Scalable Image CAT is implemented using MPEG-21 DIA's BSD tools, allowing resolution and quality reduction of JPEG2000 images.

Integration of the Scalable Video CAT will will start after the under development aceMedia Scalable Video[7] format is fixed.

3.3 Semantic Driven Real-Time CATs

Currently there are two Semantic driven real-time CATs, working respectively over images and compressed video. It intends to provide content adaptation according to semantic features extracted in real time from the incoming content.

The Flesh Colour Image CAT performs a definition of regions of interest in the original image looking for connected flesh colour zones. The resulting regions of interest are filtered depending on their shape, size or proximity to other flesh colour zones in the image eliminating the less interesting regions and providing a score for each resulting region of interest. From this point it is possible to encode the image in JPEG2000 format giving higher quality to the resulting set of regions of interest or just encoding a portion of the original image containing the region of interest with higher score (or a group of top scored regions of interest).

The Moving Regions Video CAT operates with compressed video. Currently shot segmentation followed by keyframe detection and moving objects detection is done, in order to adapt, respectively, temporally (temporal segments with low motion or reduced number of changes) and spatially (spatial ROIs ordered by motion activity). In order to perform the required content analysis in real-time, this module works over MPEG compressed domain parameters, such as DCT coefficients and coding motion vectors [8].

3.4 Transmoding CATs

Currently, only simple audiovisual to audio (or video) transmoding in available in the system, and this functionality is available via the Audiovisual Transcoding CAT. In the near future we are going to integrate real transmoding CATs: a Media2Text CAT based on the available MPEG-7 textual and Spoken Content description, and a Video2Slideshow CAT[9].

4 Conclusions

CAIN is a Content Adaptation Module targeted to the integration of different content adaptation approaches. The content adaptation starts with a common phase where the adaptation parameters are selected taking into account the media description and context description inputs. Afterwards, the most appropriate content adaptation tool is selected, taking into account their provided functionalities (adaptation capabilities, input and output media formats, …). Key-technologies in the operation of CAIN are MPEG-7 and MPEG-21, besides the coding formats (mainly, MPEG and JPEG families), and the integrated Content Adaptation Tools.

The different Content Adaptation Tools already integrated within CAIN (and some under development and integration) have been presented in this paper, showing that the different adaptation approaches allows more flexibility depending on the usage scenario.

Acknowledgements

Work partially supported by the European Commission under the 6th Framework Program, project FP6-001765: Integrating Knowledge, Semantics and Content for User-centred Intelligent Media Services (aceMedia). This work is also supported by the Ministerio de Ciencia y Tecnología of the Spanish Government under project TIC2002-03692 (DYMAS).

References

1. Vetro A., Christopoulos C., Ebrahimi T. (eds.): Universal Multimedia Access (special issue). IEEE Signal Processing Magazine 20 ,2 (2003)
2. Pereira F., Burnett I.: Universal Multimedia Experiences for Tomorrow. IEEE Signal Processing Magazine 20, 2 (2003) 63-73
3. Martínez J.M., Valdés V., Bescós J., Herranz L.: Introducing CAIN: a Metadata-driven Content Adaptation Manager Integrating Hetereogeneous Content Adaptation Tools. In: Proceedings of the WIAMIS'2005 (2005)

4. Vetro A.: Transcoding, Scalable Coding and Standardized Metadata. In: Garcia N., Martínez J.M., Salgado L. (eds.): Visual Content Processing and Representation. Lecture Notes in Computer Science,Vol. 2849, Springer-Verlag, Berlin Heidelberg New York (2003) 15-16
5. Komptasaris I., Avrithis Y., Hobson P., Strintzis M.G.: Integrating Knowledge, Semantics and Content for User-centred Intelligent Media Services: the aceMedia Project. In: Proc. of WIAMIS'2004 (2004)
6. http://ffmpeg.sourceforge.net/
7. Sprljan N., Mrak M., Abhayarathe G.C.K., Izquierdo E.: A Scalable Coding Framework for efficient Video Adaptation. In: Proc. of WIAMIS'2005 (2005)
8. Bescós J., Herranz L.: Reliability Based Optical Flow Estimation from MPEG Compressed Data. In: Proc. of VLBV05 (2005)
9. Padilla M., Martínez J.M., Herranz L.: Video Summaries Generation and Access via Personalized Delivery of Multimedia Presentations Adapted to Service and Terminal. International Journal of Intelligent Systems (2006) in press

Extrapolating Side Information for Low-Delay Pixel-Domain Distributed Video Coding

Luís Natário[1,⋆], Catarina Brites[1], João Ascenso[2], and Fernando Pereira[1]

[1] Instituto Superior Técnico, Lisboa, Portugal
luis.natario@lx.it.pt
http://www.img.lx.it.pt
[2] ISEL, Lisboa, Portugal

Abstract. Distributed Video Coding (DVC) is a new video coding approach based on the Wyner-Ziv theorem. Unlike most of the existing video codecs, each frame is encoded separately (either as a key-frame or a Wyner-Ziv frame) which results in a simpler and lighter encoder since complex operations like motion estimation are not performed. The previously decoded frames are used at the decoder to estimate the Wyner-Ziv frames - the frames are coded independently but jointly decoded. To have a low-delay codec, the side information frames (estimation of the Wyner-Ziv frames to be decoded) must be extrapolated from past frames. This paper proposes a robust extrapolation module to generate the side information based on motion field smoothening to provide improved performance in the context of a low-delay pixel-domain DVC codec.

Keywords: *distributed video coding, side information, motion extrapolation, low-delay.*

1 Introduction

Most of the existing coding schemes, namely the popular MPEG standards, are based in an architecture where the encoder is typically much more complex than the decoder mainly due to the computationally consuming operation of motion estimation done at the encoder. The Distributed Video Coding (DVC) approach based on the Wyner-Ziv (WZ) theorem [1] (which is the extension of the Slepian-Wolf theorem [2] for the lossy case with side information available at the decoder) allows reversing this scenario by shifting the motion estimation complexity from the encoder to the decoder enabling applications where the encoder's low complexity is a requirement. The Slepian-Wolf theorem states that is possible to compress in a distributed way (separate encoding and joint decoding) two statistically dependent signals at a rate similar to the rate obtained using a system where the signals are encoded and decoded jointly (as in the traditional video coding schemes). In DVC schemes, each frame is encoded independently

[⋆] The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org).

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 16–21, 2006.
© Springer-Verlag Berlin Heidelberg 2006

from previous and subsequent frames which results in a decrease of the typical encoding complexity. In order to have a low-delay codec, the frames must be decoded regardless of future frames, i.e. the side information must be created by extrapolation (as opposed to create the side information by interpolation using also future frames). The main novelty of this paper is the side information extrapolation module that is able to generate accurate side information by employing an extrapolation model that uses overlapped motion estimation, motion field smoothening and spatial-interpolation for uncovered areas. This approach enables a low-delay DVC architecture which is particularly well suited for emerging applications where the encoder complexity must be as low as possible and low-delay is a ’must have’ like in wireless low-power surveillance and mobile camera phones among others.

2 Pixel-Domain Wyner-Ziv Codec Architecture

The IST-Wyner-Ziv (IST-WZ) codec developed at IST [3] is based on the pixel-domain coding architecture proposed in [4]. The scheme, modified to support the low-delay extrapolation module, is depicted in Fig. 1.

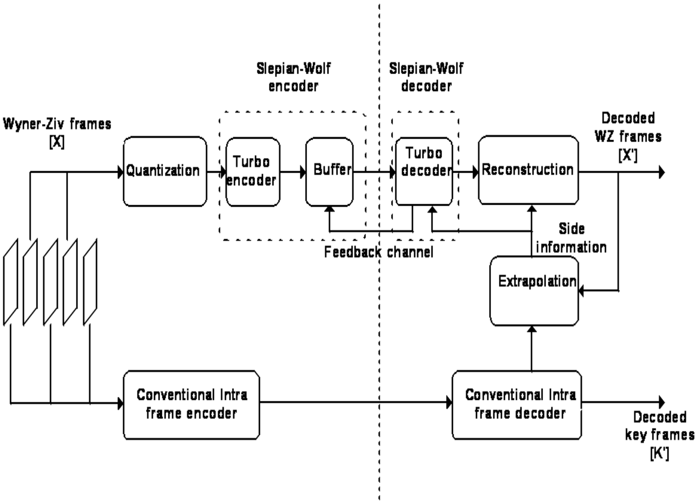


Fig. 1. Wyner-Ziv codec architecture with side information extrapolation

In the IST-WZ codec, the frames encoded are of two types: key-frames and WZ-frames. The key-frames are intra coded using H.263+, for example. The WZ-frames, after being uniformly quantized, are encoded using a turbo-based Slepian-Wolf encoder. The key-frames (and previously decoded WZ-frames, if decided) are used by the decoder to generate, by extrapolation, the side information that along with the WZ-bits received will be used to decode the WZ-frames. The Slepian-Wolf encoder generates sequences of parity bits for each bitplane output by the quantizer. These bits (which depend on the turbo encoder rate) are punctured - i.e. divided into subsets and reordered according to a given pattern - and finally stored at the Slepian-Wolf encoder's buffer. Depending on the quality of the side information generated, more or less WZ-bits will be requested (via feedback channel) by the decoder to ensure that a given WZ-frame is successfully decoded with a given bit error probability. The side information generated must be as close to the original as possible to ensure that a minimum amount of bits is requested to decode a given WZ-frame.

3 Extrapolating the Side Information

Given a video sequence, one can predict a forthcoming frame at the decoder based on the past, i.e. using the previously decoded frames, by extrapolation. The advantage of using extrapolation (and not interpolation) for the generation of the side information is to enable a low-delay codec, since to decode a given frame no future frame is needed. The performance of this type of coding architecture (DVC) is fundamentally determined by the quality of the predicted frame (side information produced at the decoder) because, if the extrapolated side information frame is very similar to the WZ-frame being decoded, few coding errors have to be corrected and, therefore, few parity bits from the encoder's buffer are requested by the decoder, resulting in a low bitrate for the WZ-frames. Several methods can be thought to extrapolate a side information frame. The simplest approach is to use the previous decoded frame. However, and since motion is generally present in video sequences, an extrapolation based on the motion observed in the previously decoded frames fits better the purpose of producing an extrapolated frame similar to the WZ-frame being decoded. The motion estimation must be done carefully in order to ensure that only true motion is captured and that a reliable extrapolated frame is produced. Since the motion observed in the previously decoded frames is projected to the WZ-frame time slot, the motion estimation targets capturing the blocks that minimize some distortion measure and are associated to the true motion. The side information extrapolation module proposed in this paper is composed of several blocks (as depicted in Fig. 2):

- Motion estimation - Motion vectors are estimated for overlapped 8×8 pixel blocks using the two previously decoded frames (Fig. 3). The block overlapping (superposition of the blocks used to perform motion estimation) is used to reduce the block artifacts in the side information frame caused by block motion-compensation.

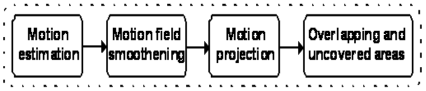


Fig. 2. Side information extrapolation module

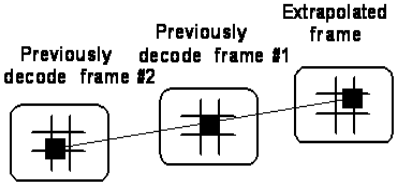


Fig. 3. Motion projection

- Motion field smoothening - For each block, a new motion vector is calculated by averaging all neighboring motion vectors. This leads to a smoothened motion vector field where true motion is captured and a better side information frame obtained.
- Motion projection - The pixels from the last decoded frame (or other) are projected to the next time instant using the motion field obtained above assuming that the motion is linear and that, therefore, the warping of frame $i-2$ into frame $i-1$ will linearly continue from frame $i-1$ to frame i (Fig. 3).
- Overlapping and uncovered areas - Whenever one pixel is estimated by more than one pixel in the previous frames, an average between the values is taken as the prediction for that position. Whenever no pixel in the previous frames is assigned as a prediction for a given frame, it is predicted by local spatial interpolation from three neighbors (up, left and up-left), scanning the frame from top to bottom and left to right.

4 Tests and Results

The tests performed envisioned evaluating, first of all, the extrapolation module with the tools proposed versus similar systems present in the literature [5]. The results obtained for the sequence *Foreman* using the first 100 frames (using a frame structure *I-WZ-I-WZ* and accounting only the WZ-frames bitrate) shows,

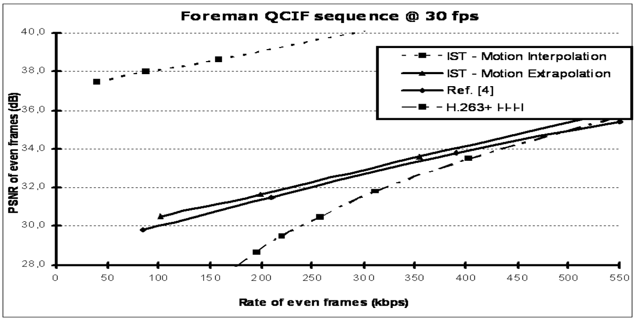


Fig. 4. RD performance comparison for Foreman

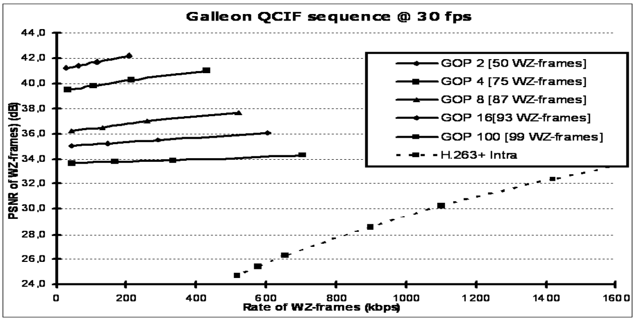


Fig. 5. RD performance for Galleon using different GOP sizes

in Fig. 4, that the system proposed here is performing slightly better than the best comparable results from the literature [5] (for the single sequence for which comparable results are available). Fig. 4 also shows that using the proposed extrapolation module either than the interpolation module proposed in [3] results in a significantly poorer RD performance. When the side information is extrapolated instead of being interpolated between two key-frames a worse prediction of the side information is obtained - this is a direct consequence of the low-delay constraint.

Fig. 5 presents the RD performance using the extrapolation module proposed, for different GOP-sizes (QCIF), for the rather still surveillance like sequence *Galleon* (src20) from the VQEG test sequence set. The usage of fewer key-frames results in a progressive decrease of the quality since that the side information generated by projecting a decoded WZ-frame does not offer the same quality of projecting a decoded intra frame. Nevertheless, the RD results obtained for all GOP sizes tested clearly outperform the RD results obtained using a H.263+ intra encoder.

5 Conclusions

The proposed Wyner-Ziv codec solution shows that it is possible to have a low-delay codec with a simple and lightweight encoder providing substantially better quality than traditional intra encoders. This extrapolation-based DVC codec is especially well suited for stable sequences (e.g. surveillance sequences) where low-delay and low encoder complexity are strong demands.

References

1. Wyner, A., Ziv, J.: The Rate-Distortion Function for Source Coding with Side Information at the Decoder. IEEE Trans. on Information Theory **22** (1976)
2. Slepian, J., Wolf, J.: Noiseless Coding of Correlated Information Sources. IEEE Trans. on Information Theory **19** (1973)

3. Ascenso, J., Brites, C., Pereira, F.: Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding. 5th EURASIP Conf. on Speech and Image Processing, Multimedia Communications and Services, Smolenice, Slovak Republic (2005)
4. Aaron, A., Zhang, R., Girod, B.: Wyner-Ziv Coding for Motion Video. Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, USA (2002)
5. Aaron, A., Rane, S., Setton, E., Girod, B.: Transform-Domain Wyner-Ziv Codec for Video. Proc. Visual Communications and Image Processing (2004) 520-528

Transmission-Efficient Image-Based Authentication for Mobile Devices

Tatiana Onali and Giaime Ginesu

DIEE, Department of Electrical and Electronic Engineering,
University of Cagliari, Piazza D'Armi, Cagliari 09123, Italy
{tatiana.onali, g.ginesu}@diee.unica.it

Abstract. A novel method for image-based authentication (IBA) in a mobile device framework is proposed and evaluated. Often, the users desire for a reliable and user-friendly authentication makes systems vulnerable and hinders the achievement of World Wide Web e-services. In particular, the development of Mobile Internet requires the definition of a user authentication system that satisfies all the requirements of security and usability and does not requires specialized hardware. If on the one hand graphical passwords are easier to recall than traditional alphanumeric passwords, on the other hand the wireless environment imposes severe constraints on data transmission capability and user friendliness, which are satisfied in the system devised.

1 Introduction

The development of wireless technology allowed for easy Internet access through mobile devices. E-commerce, home banking and online trading are just a few possibilities of Mobile Internet, while mobile phones already substitute credit cards for small payments. With the growth of such applications the problem of access security becomes increasingly important. A robust control access system is an essential condition to back the thriving of new web services, guarantying a high security level without compromising simplicity and efficiency of authentication process.

Nowadays, authentication is often performed through weak and impractical methods [2]. Most security systems employ memory-based techniques, requiring the user to precisely recall complex alphanumeric passwords. Besides, input-interface limitations represent a great constraint to user friendliness in wireless environment. Consequently, users widely choose weak passwords, as common words or short PINs, exposing the system to security threats. Furthermore, more advanced authentication systems, as token-based or biometry-based techniques, require specialized hardware, which may be very expensive or incompatible for wireless technologies.

Image-based authentication (IBA) is a valid alternative for achieving a good trade-off between security and usability in every user device, from computer terminals to mobile phones. The prerogative of a visual login system is that of providing both a good level of security, user friendliness and ease of memorization. Several experiments of cognitive science show, in fact, that pictures are easier to recall than alphanumeric passwords [5, 6, 9].

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 22 – 28, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Some graphical password systems have already been proposed for wireless applications. Awase-E [8] consists of multi-step verification stages, each requiring the identification of a pass-image among a set of 9 images. Passface [4] is a *cognometric* method of personal authentication; it is similar to Awase-E and is based on the human ability to recognize familiar faces. PassPic [7] consists in a single verification stage, requiring the selection of a correct pass-images sequence out of 9 images. Although these systems are suited for mobile terminals, they do not offer a good solution in respect of usability and data transfer. PassPic requires to remember a combination of pictures in precise order, which becomes harder to remember than alphanumeric passwords, thus nullifying the simplification introduced by the visual approach. Awase-E and Passface, instead, involve the transmission of several pictures, which is inconvenient due to bandwidth limitation of wireless channels. GPRS network providers, for instance, generally allow for a bandwidth smaller than 56kbps, while the billing system is often traffic-dependant. Besides, Awase-E and Passface offer a security level comparable to PIN codes, which is inadequate for current applications.

This paper proposes a novel image-based authentication method tailored for mobile devices, as it is relatively simple but effective, providing an adequate level of security. The proposed framework makes extensive use of the JPEG and JPEG2000 standards both for image storage and processing.

2 IBA Method

The proposed IBA method is based on a client-server interface to optimise processing, minimize data transmission and split password into several steps to improve security. The core algorithm consists of a progressive zooming technique and provides two levels of security: recognizing a pass-image and selecting a secret portion through a sequence of zoom levels.

The authentication framework provides two classical phases: registration and authentication.

2.1 Registration of Personal Data

The registration phase allows the user to acquire a personal software key, to choose the desired images for authentication and to define his graphical password. While authentication may be performed from any device, such as computer terminals, PDAs and mobile phones, registration has to be carried out from a computer terminal.

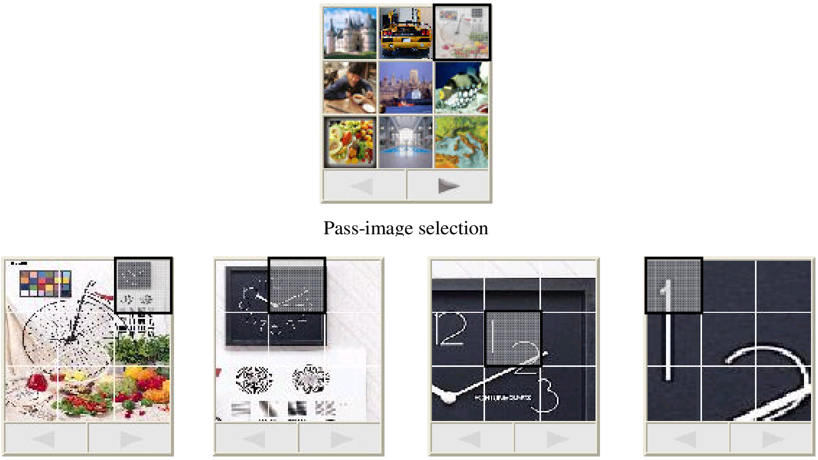
In order to register a new user the server presents a traditional form for submitting the user's personal data and devices characteristics; it generates a software key, which will be used to identify the client each time he tries to log in, customizing the authentication procedure. In particular, personal device/card codes (IMEI, SIM) may be used to allow for the unique identification of the user. In this way, the proposed method provides a two-factor authentication, based on something the user has, *i.e.*, the mobile phone, and something the user knows, *i.e.*, the password.

Subsequently, the server shows a large set of images, randomly selected from a database of JPEG and JPEG2000 images. These images should be inspired by some different themes, excluding random-art and abstract images, which might be hard to remember, compromising the usability of the proposed method [1]. The user must choose a pass-image from the visual database and a pass-detail from the selected image. Upload of personal images is allowed, although it is generally discouraged, since the authentication process may be easily guessed from personal data. As the registration process may be time consuming and requires the exchange of personal data, it is done online from a computer terminal.

2.2 User Authentication

During the authentication phase, the server manages the preliminary user and user's device identification by detecting and decrypting the software key. The visual password codes are transmitted step by step, in order to minimize the risk of sniffing. Whenever the server detects an authentication failure, the authentication process is not interrupted until the last step. Only then, the user is rejected and a notification policy is adopted.

An example of the authentication process is shown in Fig. 1. During each authentication session, the server shows up to *N* grids containing 9 images each, easily selectable using the numeric keypad. The images are arranged in random order by the server. Then, the risk of back-shoulder attack is minimized and the process of authentication might be completed before the transmission of all *N* grids. After the pass-image selection, the user has to iteratively select the correct image portion through *P* zoom levels. The values of *N* and *P* depend on the desired degree of security. The server replies to each user's request by providing the exact visual information so that refinement data are preferably transmitted. In order to do so, only the correct portion of information is transmitted at each step. The only processing required on the client's side is the exact resizing of the received image and the transmission of pass-coordinates.



Pass-image selection

1st zoom level

2nd zoom level

3rd zoom level

4th zoom level

Fig. 1. Example of authentication process with the proposed IBA method (N = 1, P = 4)

2.3 Image Processing and Transmission

In order to minimize data transmission algorithmic complexity on the user device, the major part of data processing is performed on the server side, which is required to store and manipulate a database of JPEG or JPEG2000 compressed images.

The proposed method requires 9×N images with a size proportional to *display-size*×9^{P-1} for each user. To optimize data transfer, these images are processed by the server before transmission. In particular, the server builds each of the N grids by composing 9 images into one fit the display size. In this way, it will send to the client N images of 15KB each average.

Moreover, the iterative zooming is managed through the transmission of refinement data only. Each zooming step is first predicted at the client side though a bilinear interpolation algorithm and the prediction error is sent by the server, as shown in Fig. 2. Although not optimal for the task, the standard JPEG and JPEG2000 compression algorithms have been used to transmit the prediction error images. Nonetheless, tests demonstrated a reduction of data transfer between 10% and 20% for a target PSNR of 35dB.

The proposed method has been tested on a database of 10 images by using the JPEG and JPEG2000 standard codecs and by setting a target PSNR of 30, 33 and 35dB (Table I).

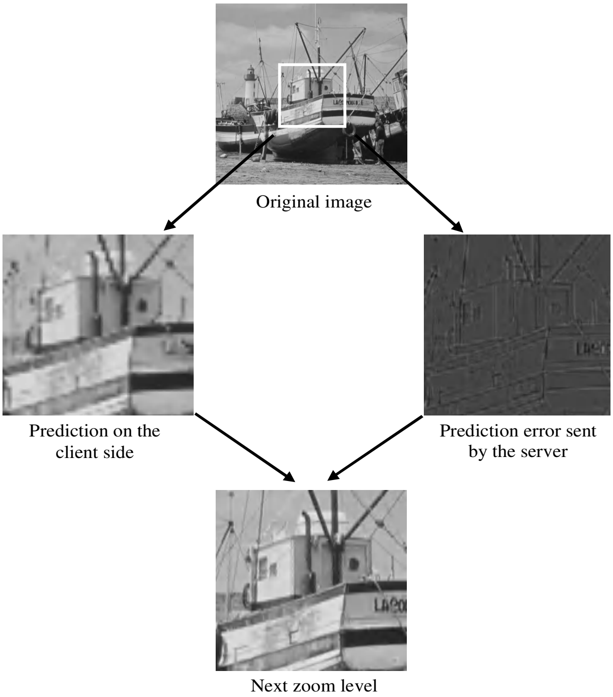

Original image
Prediction on the client side
Prediction error sent by the server
Next zoom level

Fig. 2. Example of the algorithm for iterative zooming

Table 1. Average data stream reduction through the proposed method

	JPEG		JPEG2000	
Target PSNR	Stream size (kB)	% reduction	Stream size (kB)	% reduction
30dB	5.72	38.68	2.11	59.66
33dB	11.52	21.87	6.03	30.17
35dB	16.73	11.54	9.38	16.70

Results are averaged over the dataset and show the prediction error image size and the stream reduction in respect to the original image compressed at the same target PSNR. Currently, the better rate-distortion trade-off is offered by JPEG2000. Besides, JPWL [3] defines tools and methods to implement an efficient transmission of JPEG2000 images over an error-prone wireless network.

3 Results

The proposed IBA method has been evaluated in terms of three basic requirements: data transfer, security and usability. Three state of the art graphical password systems (Awase-E, PassFace, PassPic) have been considered to compare the performance of the proposed method.

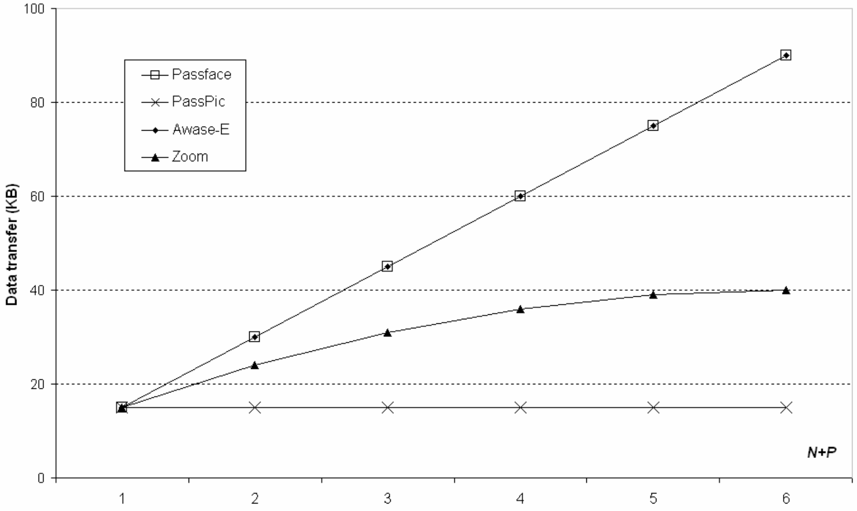


Fig. 3. Data transfer performance

The data transfer result is shown in Fig. 3. For the proposed method, the use of one image grid only ($N = 1$) is considered. In this case, the first step consists in the transmission of a composite image requiring 15KB on average. At each successive step, the size of the prediction error image decreases progressively. As a result, an average decrease of 9KB to 1KB has been recorded. On the other hand, Awase-E and Passface require one image of 15KB on average to be transmitted at each step. Finally, PassPic only requires the transmission of one image. This is the optimal choice for data transmission, but it is compromised by the need for choosing an exact image sequence with precise order, which reduces usability. The proposed method is simpler than all other visual login systems considered; it only requires the memorization of a pass-image and its pass-detail, whereas the other methods ask the user to remember at least one image for each verification stage.

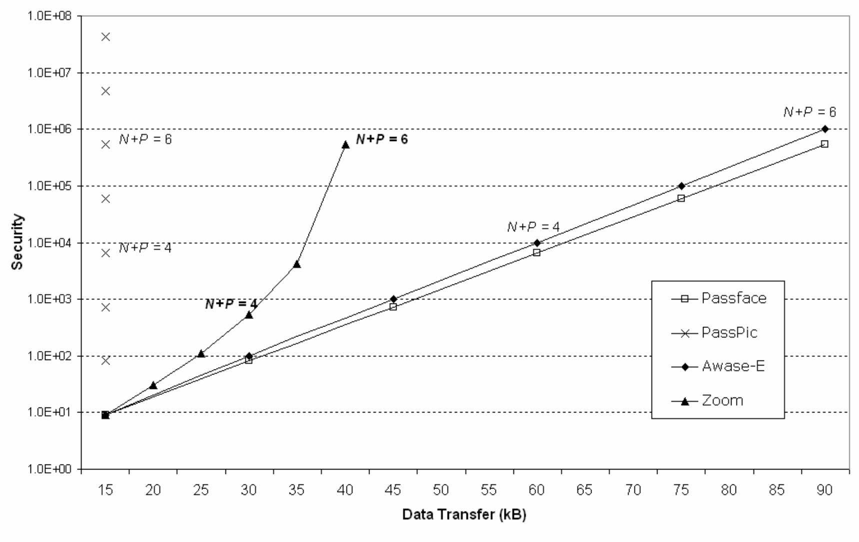


Fig. 4. Security performance

Table 2. Input combinations for the considered approaches

METHOD	SECURITY	
PassPic	9^{N+P}	N+P length of graphical password
Awase-E	9^{N+P}	N+P number of verification stages
Passface	10^{N+P}-1	N+P number of verification stages
Zoom	$N\times 9^{P+1}$	N number of first grids P number of zoom levels

Security is measured in terms of possible input combinations (Table 2) and is reported against data transfer in Fig. 4. For the proposed method the maximum number of zooming stages *P* is limited by the original image and display sizes. The proposed method greatly outperforms Awase-E and Passface. Since PassPic only requires the transmission of one image, its security/data transfer performance is superior than that of the proposed method. However, such achievement is obtained at usability expense.

4 Conclusions

Results indicate the validity of the presented method, which constitutes the better trade-off between data transfer, security and usability for user authentication on wireless portals. Furthermore, the proposed IBA method is suitable for any user device, from mobile phones to personal computers, allowing users the access to web services using the same password from every access device.

References

1. Bower, G.H., Karlin, M.B., Dueck, A.: Comprehension and memory for pictures. Memory and Cognition, Vol. 3, No. 2 (1975) 216-220
2. Cheswick, W.R., Bellovin, S.M., Rubin, A.D.: Firewalls and internet security: Repelling the wily hacker. Addison-Wesley Professional (1994)
3. JPEG 2000 image coding system – Part 11: Wireless JPEG 2000 – Working Draft version 3.1. ISO/IEC JTC 1/SC 29/WG 1 N3294
4. Passfaces™ – http://www.idarts.com/. Also in: Doi, M., Sato, K., Chihara, K.: A Robust Face Identification against Lighting Fluctuation for Lock Control. In: Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan (1998) 42-47
5. Paivio, A., Rogers, T.B., Smythe, P.C.: Why are pictures easier to recall than words?. Psychonomic Science, Vol. 11, No. 4 (1968) 137-138
6. Shepard, R. N.: Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, Vol. 6 (1967) 156–163
7. Sorensen, V.: PassPic (formerly ADS Security Wizard) – http://www.authord.com/PassPic/
8. Takada, T., Koike, H.: Awase-E: Image-based Authentication for Mobile Phones Using User's Favorite Images. Int. Symposium on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI 2003). In: Lecture Notes in Computer Science, Vol. 2795, Springer-Verlag (2003) 347-351
9. Weinshall, D., Kirkpatrick S.: Passwords you'll never forget, but can't recall. In: Proc. Conf. on Computer Human Interaction (CHI), Vienna (2004)

A Study on Visual Attack to BPCS-Steganography and Countermeasure

Michiharu Niimi[1], Hideki Noda[1], and Bruce Segee[2]

[1] Kyushu Institute of Technology, 680-4 Kawazu, Iizuka 820–8502, Japan
[2] University of Maine, 5708 Barrows Hall, Orono, ME 04469–5708, USA

Abstract. This paper discusses a visual attack to BPCS-Steganography (Bit-Plane Complexity Segmentation-Steganography) and presents a countermeasure. BPCS replaces noisy blocks with the binary patterns mapped from secret data. The logical operation called conjugate operation is applied to those if the binary patterns are simple. We need to keep the flag called a conjugation flag representing whether such conjugation had been applied to them. In the case where the conjugation flags must be embedded within cover images, the flags must be embedded into a fixed area. To this effect it is easy to recognize unnatural patterns on the LSB plane of stego-images. We propose a secure BPCS which is robust against the visual attack. It is realized by assigning the meaning of the conjugation flag to a pixel of each block and making the specific rule to embed and extract it.

1 Introduction

Steganography is the art of embedding secret information in an innocent data. Digital images are widely used as cover media because of containing much unimportant information for human visual system. To conceal the existence of embedded information, the cover images with secret data embedded must be innocent.

We have proposed a steganographic method called BPCS-Steganography (Bit-Plane Complexity Segmentation Steganography) [1, 2, 3, 4] that is an image-based steganography using image segmentation by a measure named complexity. The complexity is defined over a local region within a binary image. BPCS replaces noisy blocks on bit-planes with the binary patterns mapped from secret data. The complexity is used to determine whether blocks are noise-like. BPCS can extract embedded information by using the simple thresholding in the complexity because secret data is replaced with only complex regions. To compensate the extraction, the logical operation called conjugate operation is used for the binary patterns look like informative regions. Therefore, we need to keep, for each block, the flag called conjugation flag representing whether the conjugate operation had been applied.

In the case where we build a secure system to exchange secret messages using BPCS, which is explained latter in detail, the conjugation flags must be embedded within a fixed area of cover images. When the conjugate flags are embedded

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 29–36, 2006.
© Springer-Verlag Berlin Heidelberg 2006

in the manner, we can easily see unnatural patterns by observing the image visually. It can be used as a signature or a distinguishing mark between natural images and images with information embedded by BPCS, and it is regarded as a visual attack[5, 6, 7] to BPCS.

We propose a countermeasure to the visual attack of BPCS. The reason why BPCS is weak against the visual attack is to embed conjugation flags within a fixed area. The proposed method can embed the flags into each block by the thresholding in the complexity, as a result of the embedding, the distinguishing mark would be removed.

2 BPCS-Steganography

BPCS replaces noisy-regions on bit-planes with the secret data to be embedded. The noisy-regions are determined by the thresholding of the complexity. The complexity is based on the border-length of white and black pixels. Let P be a binary image squared $m \times m$ and k is the total length of the black-and-white border in the image. In this paper, we use four connectivity, so the total length of the black-and-white border is equal to the summation of the number of color-changes along the rows and columns in the interior of the image. Thus, the image complexity measure is defined by the following.

$$\alpha(P) = k/2m(m-1) \tag{1}$$

Where, $0 \leq \alpha \leq 1$ and $2m(m-1)$ is the length of the border on the checkerboard patterns that are the most complex pattern.

The embedding procedure in BPCS for gray-scale cover images (N bits per pixel) consists of 4 steps as follows.

Step1) The cover image in natural binary code is transformed to N-bit Gray Code. The Gray Code image is decomposed into N bit-planes, which are regarded as binary images.

Step2) The binary images are divided into small blocks which are squares of size $m \times m$. The blocks with complexities of greater than or equal to a threshold, denoted by α_{TH}, are the candidates for replacement.

Step3) Secret data can be divided into $m \times m$ bits. The $m \times m$ bits are regarded as a binary image of $m \times m$ in size. To compensate for the extraction by complexity thresholding, if the mapped binary image has complexity less than α_{TH}, the logical operation called conjugate operation using the exclusive or operation is applied to the image. Let $\alpha^*(P)$ be the complexity of the block P to which the conjugate operation is applied. We can easily calculate $\alpha^*(P)$ from its original value, that is, $\alpha^*(P) = 1 - \alpha(P)$.

Step4) We replace noisy blocks with binary images produced by the above step. We need to keep, for each block, track of whether the conjugate operation had been applied. We call it *a conjugation flag*.

Embedded information is extracted by the reverse steps of the embedding procedure by using the α_{TH} and the block size used in embedding and the conjugation flags produced after the embedding.

In BPCS, the conjugate operation is applied to the binary patterns whose complexity is less than α_{TH}. This operation makes the complexity change from $\alpha(P)$ to $1-\alpha(P)$. Thus, in the image containing secret data, the conjugation flags are useful only for the blocks whose complexity is greater than $1-\alpha_{TH}$ to distinguish whether the conjugate operation is needed to recover embedded information. In other words, the conjugation flags are not needed for the other blocks.

3 Finding Signature of Information Embedding by Visual Observation

In the secret key cryptosystem consisting of a sender and a receiver connected with an insecure channel, the important messages between them is encoded with a key at the sender and the encoded one is transmitted to the receiver over the insecure channel. At the receiver side, the received message is decoded with the key. It is possible to encode and decode with only one key. If we exchange somehow the key by using secure channel once, after that, we can keep exchanging encoded messages without exchanging the key again. If we require the convenience of the cryptosystem to BPCS, it is desirable to choose information which does not depend on image data and secret data as the key.

In order to extract the embedded information from stego-images with secret data embedded by BPCS, we need to know the information about the threshold and the block size used in embedding, and conjugation flags. Conjugation flags are not candidate for the key because those depend on image data, block size, secret data and the threshold. Therefore, conjugation flags must be embedded within cover images in the system.

The conjugation flags, however, are not embedded by the complexity thresholding because the additional conjugation flag is needed for the blocks in which the conjugations flags are embedded. Therefore the conjugation flags must be embedded into a fixed area of cover images.

Fig. 1. Gray scale representation

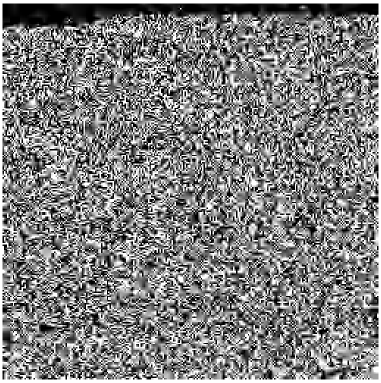


Fig. 2. LSB of Fig. 1 with Gray code

There are several ways to embed conjugation flags within a fixed area on cover images. The simplest way to do that is to embed those information into a fixed area on LSB (Least Significant Bit) plane. We can embed it without degrading the quality of cover images because it is not noticeable intensity changes in the LSB plane for gray scale images. However, we can detect unnatural patterns by looking the LSB plane. For example, Fig. 1 shows the stego-image with secret data embedded by BPCS with $m = 8$ and $\alpha_{TH} = 46/112$. The amount of the noisy region was about 31% of the original image. All of the regions are substituted with noisy patterns and the conjugation flags are embedded from the left upper block with raster scanning without checking whether blocks are noise-like. Fig. 2 shows the LSB plane of Fig. 1. We can easily see the unnatural pattern on upper part by comparing the image with the gray scale representation (Fig. 1). In general, the LSB plane for natural images looks like noisy patterns. Because we know such the characteristic of the LSB plane, it is easy to detect unnatural patterns from the stego-image, and it can be regarded as the signature of BPCS.

4 Robust Embedding Against the Visual Attack

4.1 Outline

The reason why BPCS is weak against the visual attack is to embed conjugation flags without using the complexity thresholding. Therefore, if both secret data and the conjugation flags can be embedded by the complexity thresholding, BPCS would be robust against it. To realize this idea, we embed conjugation flags into each block.

4.2 Embedding

Let S be a $m \times m$ squared binary image mapped from the bit sequences of secret data. The conjugation flag, which takes on "0" or "1", of S can be embedded within S if we make the value of a pixel on S correspond to the flag. We call the pixel assigned the flag *a control pixel*. The value of the control pixel represents that weather the conjugation operation had been applied. In this paper we define that "1" of control pixel means the conjugate operation had been applied to S.

First, the proposed method initializes the control pixel as "0", then make the bit sequences of secret data map on S except for the control pixel, next, calculates the complexity. By the complexity, the method determines whether the value of the control pixel should be changed form "0" to "1." In the following, $S_{CP=0}$ and $S_{CP=1}$ represent that the value of the control pixel is equal to "0" and "1."

We use one of the four corner of S as the control pixel. Because we consider 4 connectivity of pixels through this paper, there are 2 adjacent pixels of the control pixel. Thus, we can categorize adjacent-pixel patters into 3 patterns : the both values are 0, the both values are 1, and one is 0 and another is 1, and vice versa. Each pattern is denoted by BB, WW and BW in the following.

As mentioned earlier section, conjugation flags are not needed for the S whose complexity is within α_{TH} and $1-\alpha_{TH}$. In that case, it is better to embed secret data to the control pixel to increase data hiding capacity. In the extraction procedure, it is desirable that the meaning of the control pixel is determined by the complexity of a block in which the control pixel is, that is, when the complexity is greater than or equal to α_{TH} and less than or equal to $1-\alpha_{TH}$, the control pixel represents secret data and, otherwise, that represents conjugation flag. We can easily assigning the meaning to the control pixel in embedding, however, in some case, the meaning may change in extraction because the complexity would be affected by changing the value of the control pixel.

We show here an example of that case. Let $\beta(S)$ be $\alpha(S) \times 2m(m-1)$, β_{TH} be $\alpha_{TH} \times 2m(m-1)$, β^{*}_{TH} be $(1-\alpha_{TH}) \times 2m(m-1)$. We now consider the case where $\beta(S_{CP=0}) = \beta_{TH} - 2$ and the adjacent pixel pattern is BB. For the S, we need to apply the conjugate operation because the complexity of S is less than α_{TH}. If we make the value of the control pixel of S change to "1," then the maximum change in the border length is 2, thus the complexity of $S_{CP=1}$ may become $\lceil \beta_{TH} \rceil$. The control pixel of the $S_{CP=1}$ can be extracted as the secret data because, after the conjugate operation, $\alpha(S_{CP=1})$ falls within a range of α_{TH} and $1-\alpha_{TH}$. Thus, it would be impossible for this example to recover embedded information because the meaning of the control pixel is determined as not conjugation flag, but one bit of secret data.

To avoid this miss determination, we assign another function to adjust the complexity to the control pixel. Suppose that *CF*, *SD* and *CAB* represent conjugation flag, secret data and the complexity-adjustment bit, respectively. The control pixel holds one of CF, SD or CAB, and takes on one of 0 and 1 value. Those meaning and values depend on the complexity of S, one bit of the secret data which is embedded into the control pixel of S and the adjacent-pixel pattern of the control pixel.

Table 1. Meaning of the control pixel and its value

$\beta(S_{CP=0})$	Adjacent pixel patterns	Control Pixel	
		Meaning	Value
$\lceil\beta_{TH}\rceil - 2$ or $\lceil\beta_{TH}\rceil - 1$	BB	CAB	1
	WW or BW	CF	1
$\lceil\beta_{TH}\rceil$ or $\lceil\beta_{TH}\rceil + 1$	BB or WW	SD	b_n
	BW	SD	0 (if b_n=0)
		CF	1 (if b_n=1)
$\lfloor\beta^*_{TH}\rfloor - 1$ or $\lfloor\beta^*_{TH}\rfloor$	BB	CAB	0
	WW or BW	SD	b_n
$\lfloor\beta^*_{TH}\rfloor + 1$ or $\lfloor\beta^*_{TH}\rfloor + 2$	BB or WW	CF	0
	BW	CF	0 (if b_n=0)
		SD	1 (if b_n=1)

The proposed method, which is robust against the visual attack, is given as the following.

Step 1) Chose a pixel from four corner of S as the control pixel, and initialize it as "0."

Step 2) Map the bit sequences of secret data to the pixels of S except for the control pixel.

Step 3) Calculate the complexity of $S_{CP=0}$.

Step 4) By the calculated complexity, define the meaning and the value of the control pixel.

- If $\beta(S_{CP=0}) < \beta_{TH} - 2$, then the meaning of the control pixel of S is CF and its value is "1." Then the conjugate operation is applied to S.
- If $\beta_{TH} + 2 \leq \beta(S_{CP=0}) \leq \beta^*_{TH} - 2$, then the meaning of the control pixel of S is SD and its value is one bit of the secret data.
- If $\beta(S_{CP=0}) \geq \beta^*_{TH} + 2$, the meaning of the control pixel of S is CF and its value is "0."
- Otherwise the meaning and the value of the control pixel depend on the complexity, the adjacent pixel patterns and a bit to be embedded into the control pixel. Table 1 shows the meaning and the value in each case. In the table, b_n means a bit of secret data which is embedded into the control pixel.

Step5) Replace noisy blocks with S.

4.3 Extraction

Following BPCS extraction steps, we can extract the blocks, which are denoted by S', containing the secret data by the complexity thresholding. All of S' can be satisfied the following inequality about their complexity.

$$\alpha(S') \geq \alpha_{TH} \tag{2}$$

Basically, the embedded secret data is recovered from S' by the simple rule described below.

- When the complexity of S' is less than or equal to $1 - \alpha_{TH}$, then the control pixel means a bit of secret data. Thus all bits of S' are a part of secret data.
- When the complexity S' is greater than $1 - \alpha_{TH}$, then the control pixel means a conjugation flag. If the conjugation flag is equal to "1," then we apply the conjugate operation to it. The all bits except for the control pixel of the S' applied the conjugate operation are a part of secret data.

In order to take account of Table 1 used in the embedding procedure, in addition, we need the exception rule described bellow.

- When the complexity of S' is equal to β_{TH} or $\beta_{TH} + 1$, the control pixel means CAB if the its value is equal to "1" and the adjacent pixel pattern is BB.
- When the complexity of S' is equal to $\beta^*_{TH} - 1$ or β^*_{TH}, the control pixel means CAB if the its value is equal to "0" and the adjacent pixel pattern is BB.

Embedded secret data can be recovered after the above steps are performed for all blocks containing secret data.

5 Experimental Result and Discussions

In order to demonstrate the efficiency of the proposed method, we present an experimental result. We used the GIRL (256 × 256, 8 bit/pixel) image as cover image. Noisy-regions determined with $m = 8$ and $\alpha_{TH} = 46/112$ were replaced with another noisy pattern. We are unable to detect unnatural pattern from Fig. 3. in other words, we are unable to distinguish between the LSB of the cover image and that of the stego-image.

The proposed method decreases 1 bit of hiding data capacity for blocks whose complexity is less than α_{TH} or greater than $1 - \alpha_{TH}$. If α_{TH} is getting close to 0, then the number of blocks we need to apply conjugate operation decreases.

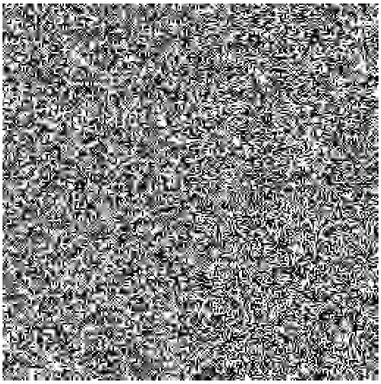


Fig. 3. LSB plane of the image with secret data embedded by the proposed method

In such cases, the proposed method is useless because conjugation flags are not needed in extracting embedded information. However, image degradation will occur on the image data as detectable patterns. Moreover, by the comparison between bit-planes, one can easily recognize the existence of embedded information. Therefore, the proposed method is very effective, that is secure, when we embed secret data with keeping high image quality by using BPCS.

6 Conclusion

We have presented an improved BPCS-Steganography that removes an identifying signature that can be found in conventional BPCS-Steganography. The proposed method can embed both secret data and the conjugation flags into cover images with image segmentation using the complexity thresholding, therefore, unnatural patterns which can be used as the signature do not appear in the stego-images.

References

1. Niimi, M., Noda, H., Kawaguchi, E.: Steganography based on region segmentation with a complexity measure. Systems and Computers in Japan **30**(3) (1999) 1–9
2. Niimi, M., Eason, R., Noda, H., Kawaguchi, E.: Bpcs-steganography to palette-based images using luminace quasi-preserving color quantization. IEICE Trans. on Fundamentals **Vol.E85-A**(9) (2002) 2141–2148
3. Ouellette, R., Noda, H., Niimi, M., Kawaguchi, E.: Topological ordered color table for bpcs-steganography using indexed color images. IPSJ Journal **42**(1) (2001) 110–113
4. Noda, H., J. Spaulding, M.S., Kawaguchi, E.: Application of bit-plane decomposition steganography to jpeg2000 encoded images. IEEE Signal Processing Letters **9**(12) (2002) 410–413
5. Wayner, P. In: : “Disappearing Cryptography second edition:Information Hiding: Steganography & Watermarking. Morgan Kaufmann Publishers (2002) 303–314
6. Katzenbeisser, S., Petitcolas, F.A. In: Information Hiding : Techniques for Steganography and Digital Watermarking. Artch House (2000) 79–93
7. N. F. Johnson, Z.D., Jajodia, S. In: Information Hiding : Steganography and Watermarking - Attacks and Countermeasures. Kluwer Academic Publishers (2001) 47–76

Fast Intra 4X4 Mode Elimination Approaches for H.264

Dajun Wu[1], Keng Pang Lim[1], Si Wu[1], Zhicheng Zhou[1], and Chi Chung Ko[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[2] Department of ECE, National University of Singapore, Singapore 117576
djwu@i2r.a-star.edu.sg

Abstract. This paper presents two fast Intra 4x4 mode elimination approaches for H.264. They terminate at an early stage during the Intra 4x4 mode decision process. The lossless approach checks cost after each 4x4 block Intra mode decision, and jumps directly to Intra 16x16 mode decision if the cost is higher than minimum cost of Inter mode. The lossy approach shortens Intra 4x4 mode decision process by using some low-cost preprocessing to make prediction, and checking if the cost is higher than some fraction of min cost of Inter mode. Experimental results show that the lossless approach can reduce the encoding time by 7.5% on average. The lossy approach can reduce encoding time by 13.87% on average with average PSNR loss of 0.048 db and average bit rate increment of 0.76%.

1 Introduction

Recently, a new video coding standard, known as H.264, has been enacted jointly by ISO and ITU [1]. It outperforms all the previous video coding standards in terms of bit rate and video perceptual quality. The significant improvement is acquired by adopting many advanced approaches such as directional prediction in intra block coding, variable block size motion estimation (ME), multiple reference frame ME, integer transform, in-loop deblocking filter and context-based adaptive binary arithmetic coding (CABAC). In addition, cost is used as the underlying criteria to guide the encoding process in a manner that given a QP, only some specific combination of mode, block size, directional prediction in the case of intra mode coding, motion vector and reference frame in the case of inter mode coding will be chosen because of incurring lowest cost.

However, this significant performance gain is obtained at the cost of high computational complexity, especially in the encoding processing. The new features incur more processing requirement and make the encoding complex. Every possibility of mode, block size, directions, motion vector and reference frame has to be tried in an exhaustive approach in order to obtain an optimal point. For instance, block intra mode could have up to nine directional predictions of which only one with the lowest cost would be selected. If rate distortion optimization (RDO) option in the encoder is switched on, obtaining rate will be more complex since in addition to transform, quantization and inverse quantization, every macroblock must finish entropy coding to get its real bit rate. As a result, fast algorithms tailored for H.264 are indispensable.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 37 – 43, 2006.
© Springer-Verlag Berlin Heidelberg 2006

In the literature, several algorithms of fast motion estimation have been proposed for H.264 [2][3]. Instead of trying every search point in the search window, they perform motion estimation using only a small portion of the search points. An effort has also been made by Pan et al to explore the fast algorithm for intra spatial prediction [4]. The local edge information is used to reduce the amount of calculations in intra prediction. Using edge direction histogram derived from the edge map, only a small number of intra prediction modes are used for RDO calculation. Therefore the speed of intra coding is greatly increased. However, the preprocessing consumes much computation power and the overall algorithm is suitable for high complexity mode of H.264, in which RDO is switched on. A recent paper [5] provides an approach to combining Inter and Intra mode decision together. The approach will skip Intra mode decision if selected Inter mode is good enough according to some criteria.

As a contrast, our approach targets at the case when RDO is off (low complexity mode of H.264) and tries to terminate the Intra 4x4 mode decision process early by using some prior preprocessing information and the minimum cost obtained during the Inter mode decision process. There are two approaches: 1) lossless approach, which instead of making judgment after performing all the 16 4x4 block Intra mode decision, checks cost after each 4x4 block Intra mode decision, and terminates if the cost is higher than minimum cost of Inter mode; 2) lossy approach, which by using some low cost preprocessing to obtain prior knowledge, terminates the whole MB Intra 4x4 mode decision if the cost is higher than some fraction of minimum cost of Inter mode. The lossless approach can reduce the encoding time by 7.5% on average and the lossy approach can reduce encoding time by 13.87% on average with average PSNR loss of 0.048 db and average bit rate increment of 0.76%.

The organization of the paper is as follows: Part 1 gives an introduction to this paper; Part 2 introduces briefly about H.264 Intra mode decision process; Part 3 presents the ideas of our fast approaches; Part 4 shows the results of two methods. Part 5 concludes the paper.

2 Overview of H.264 Intra Mode

Intra coding is used when only spatial redundancies within a video frame are exploited. Traditionally, Intra mode macroblocks are encoded by directly applying the transform, which generates much larger number of data bits compared to that of inter coding. To increase Intra coding efficiency, spatial correlation among adjacent blocks in a given frame is exploited in H.264. For example, the block of interest can be predicted from the surrounding blocks along some direction. The difference between the actual block and their prediction is then coded.

Intra coded macroblocks may use either 16x16 (Intra 16x16 mode) or 4x4 (Intra 4x4 mode) spatial prediction modes for luminance components. Four sub-modes are available with 16x16 prediction. A 16x16 macroblock can be predicted from the previously adjacent reconstructed pixels that are available due to the raster order decoding of macroblocks: vertical prediction from pixels above, horizontal prediction from pixels to the left, and plane prediction by spatial interpolation between these two sets of pixels. Nine sub-modes are available with 4x4 prediction. A 4x4 sub-block can be predicted from the previously adjacent reconstructed pixels that are available due to the raster order decoding of each 8x8 block within a macroblock, and the nested raster order decoding of each 4x4 sub-block with each 8x8 block. Due to this decoding order, not all of the 4x4 prediction modes have the decoded pixels available in their desired prediction direction. In this case, the closest available decoded pixel is used. For the chrominance (chroma) components, similar operations are done.

For each 4x4 block in Intra 4x4 mode decision, the prediction direction leading to the lowest cost is selected as best direction. Thereafter, integer transform, quantization, inverse quantization and inverse integer transform are applied in order to form the reconstructed block for future usage by neighboring blocks. The total Intra 4x4 cost for the whole macroblock is summed after processing all 16 4x4 blocks.

As an overall procedure including Intra mode decision, a typical macroblock mode decision in H.264 could be summarized as follows: 1) Perform motion estimation, get the best mode and record its minimum cost. 2) Perform Intra 4x4 mode decision for the whole macroblock and calculated the sum of minimum cost of each 4x4 block, 3) Perform Intra 16x16 mode decision for the whole macroblock and calculated its corresponding minimum cost. 4) Choose the mode having the lowest cost. The procedure is also illustrated in Fig. 1, in which the parts enclosed by dotted lines are excluded.

3 Fast Intra 4x4 Mode Elimination

3.1 Lossless Approach

This approach, instead of judging after making Intra mode decision on all the 16 4x4 blocks, checks accumulated cost after each 4x4 block intra mode decision, and

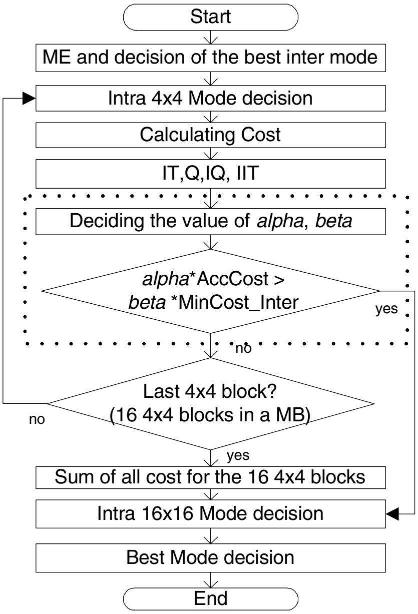


Fig. 1. Overall mode decision process

terminates the process if the cost is higher than the minimum cost of Inter mode. This early termination is very useful since the time consuming procedures of directional predictions, integer transform, quantization, inverse quantization and inverse integer transform are all skipped. As shown in Fig. 1, the blocks enclosed by dotted lines illustrate the main idea. Notice that both alpha and beta is equal to 1, and the condition is thus simplified as AccCost> MinCost_Inter, where AccCost is the accumulated cost and MinCost_Inter denotes the minimum cost of Inter mode. In this approach, there is no PSNR and/or bit rate change since this is a lossless approach.

3.2 Lossy Approach

In order to further improve the speed of encoder, we try to use some prior knowledge to guide in Intra 4x4 mode decision process. If we could predict beforehand the relative relationships of the cost for each of the 16 4x4 blocks, it is highly probable that the prediction process can be shortened. Border difference could be a good metrics to make this low cost preprocessing, since it reflects to a large extent the relative resemblance of neighboring blocks. As illustrated in Fig.2, BD (border difference) of current 4x4 block j is defined as:

$$BD(j) = \sum_{i=0}^{3} \left(\left| Y_{org}(x_j, y_j + i) - Y_{org}(x_j - 1, y_j + i) \right| + \left| Y_{org}(x_j + i, y_j) - Y_{org}(x_j + i, y_j - 1) \right| \right) \tag{1}$$

where Y_{org} is the original frame, $Y_{org}(x, y)$ represents the luminance value at position (x, y), (x_j, y_j) is the horizontal and vertical coordinates of up-leftmost point of 4x4 block j respectively, and j ranges from 1 to 16. It could be clearly shown in Fig. 2. Accumulated Border difference at block j can thus be represented as:

$$ABD(j) = \sum_{k=1}^{j} BD(k) \tag{2}$$

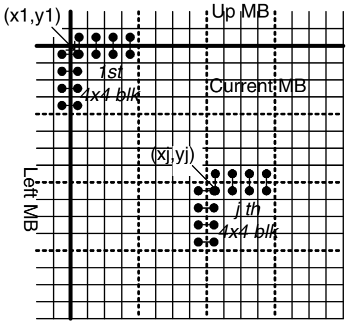


Fig. 2. Border difference of 4x4 block

Sum of border difference (SBD) is derived as:

$$SBD = \sum_{j=1}^{16} BD(j) \tag{3}$$

Border difference percentage (BDP) can thus be denoted as:

$$BDP(j) = BD(j) / SBD \tag{4}$$

Accumulated border difference percentage is:

$$ABDP(j) = \sum_{k=1}^{j} BDP(k) \tag{5}$$

Using the above information, the condition shown in Fig. 1, *alpha**AccCost > *beta**MinCost_Inter, is concreted using the following inequation:

$$1 * AccCost > ABDP(j) * MinCost\_Inter \tag{6}$$

In order to avoid division operation, the condition is further transformed as:

$$SBD * AccCost > ABD(j) * MinCost\_Inter. \tag{7}$$

Thus, *alpha* = *SBD*, and *beta* = *ABD(j)*.

In principle, if the prediction is correct, we could decide whether to continue or terminate just after the first 4x4 block is processed. But due to the reason that the prediction may not be very accurate, the accumulated cost is checked continuously and Intra 4x4 mode decision is terminated only when the condition is satisfied.

4 Experimental Results

We implemented our proposed approaches on an optimized encoder based on H.264 reference software [6]. The encoder is tailored for real-time application on normal PC

Table 1. Codec Performance (QP=28)

	Lossless		Lossy	
Sequence	TS(%)	PCH	BCH	TS(%)
Container (qcif)	6.36	-0.048	1.69	8.65
Mobile (qcif)	9.79	0.002	-0.05	14.7
Silent (qcif)	7.32	-0.051	1.58	12.49
Akiyo (cif)	4.62	-0.051	1.17	7.42
Bus (cif)	6.94	-0.044	0.42	13.99
Paris (cif)	10.79	-0.066	0.84	19.32
Tempete (cif)	8.01	-0.02	0.43	13.92
Container (cif)	7.67	-0.011	1.39	13.5
Mobile (cif)	9.96	-0.006	0.02	15.29
Waterfall (cif)	10.77	-0.011	0.63	16.03

Table 2. Codec Performance (QP=32)

Sequence	Lossless		Lossy	
	TS(%)	PCH	BCH	TS(%)
Container (qcif)	7.48	-0.035	0.07	10.86
Mobile (qcif)	9.14	0.011	0.13	13.94
Silent (qcif)	10.8	-0.186	2.08	15.8
Akiyo (cif)	6.57	-0.11	1.14	8.95
Bus (cif)	6.48	-0.063	0.79	13.48
Paris (cif)	7.17	-0.089	1.28	13.94
Tempete (cif)	6.44	-0.044	0.46	13.85
Container (cif)	6.62	-0.006	0.4	14.54
Mobile (cif)	8.77	-0.013	0.2	14.46
Waterfall (cif)	7.79	-0.007	0.03	15.3

Table 3. Codec Performance (QP=36)

Sequence	Lossless		Lossy	
	TS(%)	PCH	BCH	TS(%)
Container (qcif)	6.96	-0.014	1.6	12.66
Mobile (qcif)	7.42	-0.005	0.47	13.39
Silent (qcif)	7.9	-0.157	1.40	13.35
Akiyo (cif)	6.65	-0.058	1.36	12.65
Bus (cif)	5.33	-0.057	0.71	12.86
Paris (cif)	7.89	-0.101	1.52	15.22
Tempete (cif)	5.8	-0.067	0.74	13.86
Container (cif)	7.51	-0.06	1.06	14.96
Mobile (cif)	7.72	-0.021	0.24	14.9
Waterfall (cif)	8.37	0.003	0.33	17.37

Table 4. Codec Performance (QP=40)

Sequence	Lossless		Lossy	
	TS(%)	PCH	BCH	TS(%)
Container (qcif)	7	-0.009	0.1	13.31
Mobile (qcif)	6.25	0.001	0	13.06
Silent (qcif)	7.11	-0.094	-0.09	13.83
Akiyo (cif)	8.19	-0.067	1.73	12.83
Bus (cif)	5.04	-0.098	0.19	13.7
Paris (cif)	8.43	-0.035	0.69	16.47
Tempete (cif)	8.65	-0.065	0.68	16.51
Container (cif)	7.37	-0.088	1.65	14.81
Mobile (cif)	6.04	-0.026	0.20	14.18
Waterfall (cif)	4.9	-0.055	1.23	14.63

platform. PMVFAST motion estimation algorithm is used due to its simplicity and low complexity [7]. The test conditions are as follows: 1) MV search range is ±32 pixels; 2) Hadamard transform is used; 3) optimization is disabled; 4) reference frame number equals to 1; 5) CAVLC is enabled; 6) MV resolution is ¼ pixel; 7) GOP structure is IPPP; 8) the number of frames in a sequence is 150. A group of experiments were carried out on the test sequences with the 4 quantization parameters, i.e., QP=28, 32, 36, and 40 as specified in [8]. It should be noted that the test sequences are representative in the sense that they range from sequences of not too much motion such as ‘Akiyo’ and ‘Silent’ to sequences with much higher motion such as ‘Mobile’ and ‘Bus’. Experimental results are shown in Table 1 to Table 4, where PCH means PSNR change, BCH is abbreviation for bit rate change and TS denotes time saving. It is observed that compared to the encoder not adopting our approaches, the proposed lossless approach can reduce the encoding time by 7.5% on average. The lossy approach can reduce encoding time by 13.87% on average with average PSNR loss of 0.048 db and average bit rate increment of 0.76%.

5 Conclusion

This paper presents two fast Intra 4x4 mode elimination approaches for H.264. The lossless approach checks cost after each 4x4 block intra mode decision, and terminate if the cost is higher than min cost of Inter mode. The lossy approach, by using some low cost preprocessing to make prediction, terminate if the cost is higher than some fraction of minimum cost of Inter mode. Experimental results show that the lossless approach can reduce the encoding time by 7.5% on average. The lossy approach can reduce encoding time by 13.87% on average with average PSNR loss of 0.048 db and average bit rate increment of 0.76%.

References

1. “Information technology - Coding of audio-visual objects - Part 10: Advanced video coding,” Final Draft International Standard, ISO/IEC FDIS 14496-10
2. Li, X., Wu, G.: Fast Integer Pixel Motion Estimation. JVT-F011, 6th Meeting, Awaji Island, Japan (December 2002)
3. Chen, Z., Zhou, P., He, Y.: Fast Integer Pel and Fractional Pel Motion Estimation for JVT. JVT-F017, 6th JVT Meeting, Awaji Island, Japan (December 2002)
4. Pan, F., Lin, X., Susanto, R., Lim, K.P., Li, Z.G., Feng, G.N., Wu, D.J., Wu, S.: Fast Mode Decision Algorithm for Intra Prediction in JVT. JVT-G013, 7th JVT meeting, Pattaya (March 2003)
5. Jeon, B., Lee, J.: Fast mode decision for H.264. JVT-J033, 10th JVT Meeting, Hawaii, United States (Dec. 2003)
6. Sühring, K.: 264/AVC Software Coordination. http://iphome.hhi.de/suehring/tml/
7. Hosur, P.I., Ma, K.K.: Motion Vector Field Adaptive Fast Motion Estimation. Second International Conference on Information, Communications and Signal Processing (ICICS ’99), Singapore (Dec., 1999)
8. Sullivan, G.: Recommended Simulation Common Conditions for H.26L Coding Efficiency Experiments on Low Resolution Progressive Scan Source Material. VCEG-N81, 14th meeting: Santa Barbara, USA (Sept. 2001)

Backward-Compatible Interleaving Technique for Robust JPEG2000 Wireless Transmission

Fabrizio Frescura and Giuseppe Baruffa

Department of Electronic and Information Engineering (DIEI), University of Perugia,
Perugia 06125, Italy
{frescura, baruffa}@diei.unipg.it

Abstract. This paper presents a forward error correction technique that can be applied to protect JPEG2000 streams during their transmission over wireless channels with both sparse and packet error statistics. To this aim, we adopt the tools of the JPEG2000 Wireless (JPWL) standard, by introducing a method for Unequal Error Protection (UEP) and a virtual interleaving technique of the stream. All these techniques are introduced while keeping a complete backward-compatibility with the existing standard.

1 Introduction

The delivery of multimedia contents over IP based networks is emerging as a fundamental application for a number of source coding techniques, both in the fields of audio and/or visual content distribution. Video streaming (VS), for example, plays a major role: due to its real-time nature, video streaming typically has bandwidth, delay and loss requirements that have to be strictly satisfied [10].

The transport of video is a well known challenge for wireless IP networks. In the case of TCP, the variable delay is unacceptable for many types of real-time applications, such as video conferencing or security and surveillance.

Many video applications, also based on intra-frame source coding algorithms, as JPEG2000 [8], make now use of UDP or RTP for IP based transmission, where delay jitter is reduced at the expense of packet loss, and they are designed to tolerate it. A further way to improve video delivery performance is to provide adequate protection of video codestreams by means of a proper interleaving mechanism [2] and unequal error protection schemes [5], [9]. Interleaving is feasible and acceptable in video delivery, since it may be embedded into buffering at encoder (for rate control) and player (client) level. A buffer/interleaver of even few seconds may be tolerated by the users' experience, while it provides proper protection and error spreading capabilities.

The Wireless MAN environment (WMAN) [3] will eventually become one of the most widespread distribution systems for video delivery applications, such as for example TV broadcasting via VS. In this case, Forward Error Correction (FEC), in conjunction with inter- and intra-frame virtual data interleaving, is a method for controlling the error rate and, consequently, the propagation of errors in the stream, even in case of packet loss.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 44 – 50, 2006.
© Springer-Verlag Berlin Heidelberg 2006

2 Error Protection with the JPWL Standard Tools

The JPEG2000 standard [8] is a lossy source coding algorithm, which is currently being used even in high quality applications, such as Digital Cinema (DC) [1]. The standard defines also a set of error resilience tools (to detect and conceal erroneous data) without any correcting capability. In order to ensure a reliable transmission over wireless channels, an extension to JPEG2000 has been considered, known as JPWL (JPEG2000 for Wireless Applications) [4].

JPWL defines new codestream markers and only suggests guidelines to implement techniques for error protection, such as Unequal Error Protection (UEP), data partitioning and intelligent ARQ.

The codestream itself can be represented as a sequence of codestream packets: *H* headers and *D* data. In a JPEG2000 encoded codestream, the Minimum Square Error (MSE) can be used as an indicator for the image quality: it is a decreasing function in the number of decoded bits. In our case, we suppose that M_i is the image MSE provided that data packets *i*, *i*-1, *i*-2, …, and 1 have been correctly decoded. Moreover, we suppose that no effort is done to accept further data after that a generic packet is signaled as corrupted by unrecoverable errors. Denoting by $P^{(i)}$ the probability that error-protected packets 1 to *i* are received and decoded (after error correction), the expected average value for MSE is

$$\bar{M} = \mathrm{E}\{M_i\} = \sum_{i=0}^{D} M_i P^{(i)} \ . \tag{1}$$

It is clear that the average MSE strongly depends upon the set of the packet error probabilities, which in turn are decided by the particular Error Protection Scheme (EPS) used [9].

We consider the protection of the data codestream with the Reed-Solomon RS(*n*, *k*) block codes, as specified in the JPWL standard. In order to simplify our scheme, we will assume that the message length k_i=*k* is predefined and constant for every codeword *i* of every packet, whereas the codeword length n_i may vary from packet to packet (but it is constant inside a packet), thus giving rise to a set (a row vector with *D* entries) of codeword lengths $\mathbf{n} = (n_1, n_2, \ldots, n_D)$.

The aim of our optimization is to find an EPS which minimizes the expected value of MSE in (1), for a given Bit Error Rate (BER) on the Binary Symmetric Channel (BSC), subject to the bound that the obtained average code rate is equal to *R*. Generally, an UEP scheme results in improved final average quality over an Equal Error Protection (EEP) scheme.

In order for this method to deal with either Binary Symmetric Channels (BSC) or packet-loss channels, it is mandatory that burst errors, typical of WLAN channels, become less influent. This can be achieved, with a data interleaving strategy, both in inter-frame and intra-frame mode. Since, in this case, this technique would corrupt the syntax of the JPEG2000 codestream, we adopt a virtual interleaving scheme, where parity bytes are not computed on a consecutive stream of data, but are relevant to a periodic sampling of the stream (Fig. 1-a), i.e. the message words are composed by nonconsecutive data bytes [5].

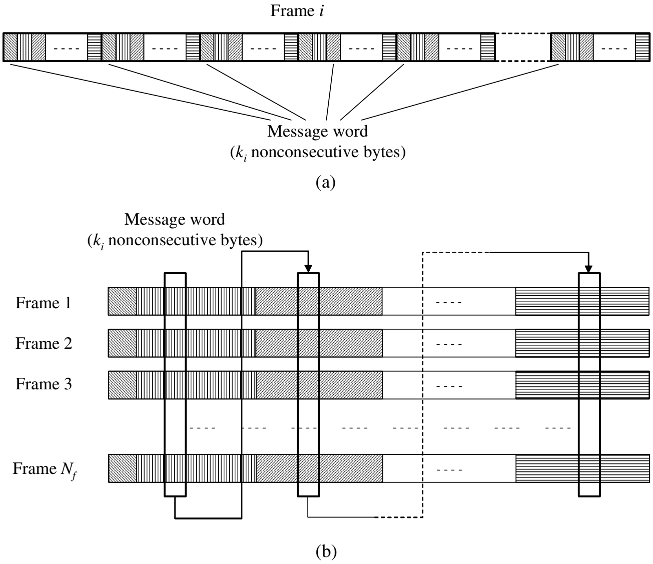


Fig. 1. In modified JPWL protection, the RS parity bytes are computed for a message word of k_i nonconsecutive data bytes, i.e. interleaved data bytes, in intra-frame mode (a). Differently, parity bytes are computed for a message word composed of bytes interspersed within N_f frames (inter-frame mode) (b). Note how, in both cases, the syntax of the JPEG2000 codestream is not modified and fully complies with the standard.

This way, even a long error burst or a network packet loss, does not necessarily imply that the whole codestream packet is lost, since each data byte is protected by different, noncontiguous, parity data blocks. Moreover, when the single frame size is small, it may happen that the packet errors imply a complete loss of the frame. In this case, we may think to extend the interleaving in time, crossing frame boundaries, and the parity bytes are computed on message bytes spread over a number N_f of compressed frames (Fig. 1-b). Note, however, that in both cases only the composition of message words upon which redundancy bytes are calculated is modified: no effective interleaving is applied on the original codestream, thus preserving its syntax.

3 Simulation Results

In order to find an optimal EPS, a simple single-frame sequence (the *Lena* image) has been analyzed. The results plotted in Fig. 2 represent the optimal allocation of error correction capability along the codestream: UEP solutions are compared to the EEP solution. It is clear that the PSNR advantage can become significant, especially for high values of the channel BER. As expected, when BER is below 10^{-3}, the two methods give similar solutions.

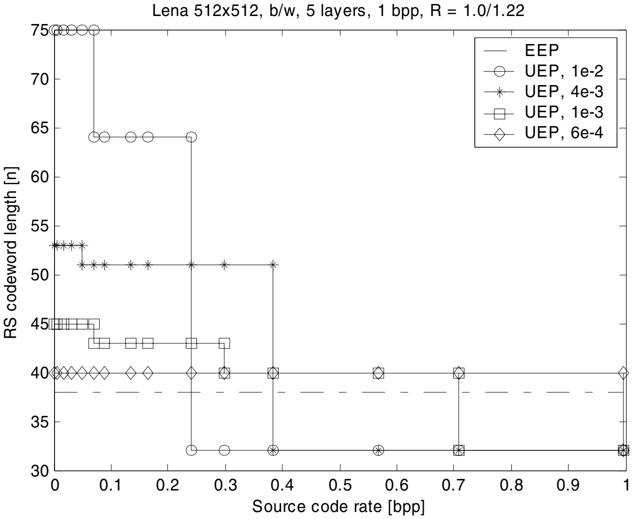


Fig. 2. Optimized error protection schemes obtained by simulation, for different values of the BSC channel target BER. The image used is Lena 512x512, b/w, 5 layers, 1 bpp.

The performance of the interleaving technique over packet-loss channels has been evaluated using a 2048x1080 (2K) image compressed according to DC specifications.

We have supposed to transmit the video stream over the wireless channel using the Motion JPEG2000 standard [6], which basically consists in a sequence of intra-coded frames plus some additional header information. The compressed sequence image data have been channel encoded according to the JPWL specifications, using the RS(37,32) error correction code. Two sets of simulations, related to different channel models, have been performed.

For the first set of simulations (Fig.3), a Gilbert-Elliot (GE) discrete channel model [7] has been adopted: this channel is characterized by the presence of error bursts, and it can also be represented by its average BER (the BER of an equivalent BSC). In this case, a number of 80 transmission and decoding trials has been simulated. Figure 3 shows the achieved PSNR performance for the transmission of a single image over a GE channel, being this equivalent to an interleaving depth of a single frame (intra-frame interleaving). As expected, the performance gain with respect to the absence of interleaving is evident. The same figure also shows the achieved performance for an interleaving depth of 2 and 6 frames (inter-frame interleaving). In this case, the higher interleaving length results in a marginal improvement in the performance gain.

As shown, the interleaving performance is substantially identical when different interleaving depths are observed. This behavior can be explained considering that the interleaving depth is already large enough to prevent the error bursts from exceeding the error correction capability of the adopted RS code, also in the single frame case.

For the second set of simulations, a different transmission channel model has been designed. It consists in partitioning the compressed video sequence into packets of fixed length, with a certain probability of packet loss. If a packet is lost, it is replaced

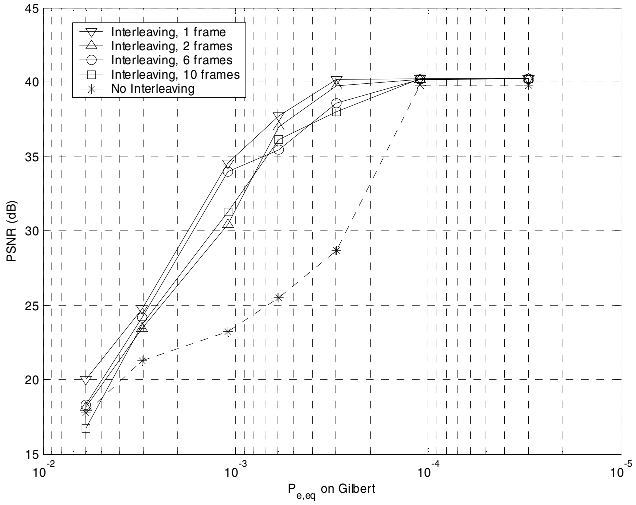


Fig. 3. Average PSNR performance for the transmission of a 2K DC sequence over a Gilbert-Elliot channel. The results for interleaving and no interleaving techniques are compared, for different values of interleaving depth.

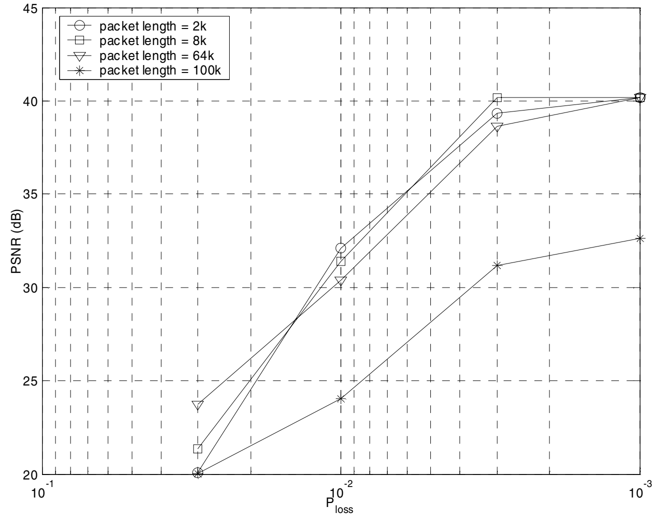


Fig. 4. Average PSNR performance for the transmission of a 2K DC sequence over a packet-loss channel. The interleaving is performed over a single frame, and the results obtained for different packet lengths are compared.

by a random sequence of bytes, thus simulating the behavior of a transmission protocol (such as, for example, UDP) over a noisy discrete channel.

For each value of the probability of packet loss (P_{loss}), a number of 400 iterations has been simulated. Figure 4 shows the obtained results. With a packet length of up to

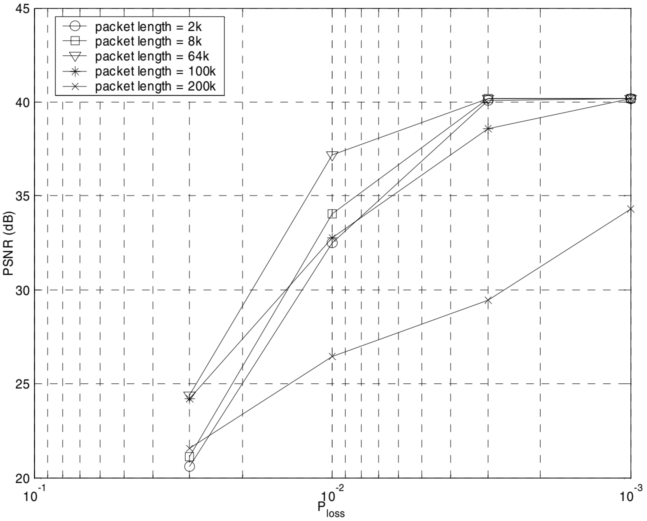


Fig. 5. Average PSNR performance for the transmission of a 2K DC sequence over a packet-loss channel. The interleaving is performed over two frames, and the results obtained for different packet lengths are compared.

64kB (65,536 bytes), the performance of the various interleaved configurations is basically the same; there is, however, a significant difference with respect to a packet length of 100kB (131,072 bytes). Since the maximum number of correctable errors is 2, for the adopted RS(37,32) code, the theoretical "breakdown" value is two times the interleaving depth, which is 81,360 bytes. This value does not represent, however, a typical case, because some of the most common protocols do not allow for such a large dimension of packets (e.g., the maximum length for a UDP packet is 64kB).

Finally, Fig. 5 shows the results for the transmission of a sequence where inter-frame interleaving is performed over 2 frames, with the same number of 400 iterations per trial. Since, in this case, the depth is doubled, the theoretical breakdown value is of 162,720 bytes.

4 Conclusion

In this paper we have presented a technique that can be adopted for the effective error protection of a JPEG2000 based video stream, when transmitted over channels with both sparse and packet error statistics. First, we introduced the JPWL standard, and then we presented a method for finding the optimal channel coding rate assignment among the header and image data, provided that a maximum average PSNR is expected on the decoded image. Moreover, virtual intra- and inter-frame interleaving can be performed on the compressed frames, still keeping complete backward compatibility with the JPEG2000 codestream syntax, and greatly improving the PSNR performance for the transmission over packet-loss channels.

Acknowledgement

Part of this work was supported by the European Union under the project "World-Screen", Grant No. 511333 (http://www.worldscreen.org).

The authors wish to thank A. De Angelis for the implementation of the C code used for the simulations.

References

1. Digital Cinema Initiatives, LLC, "Digital Cinema System Specification V5.1 (draft a0331)," April 2005.
2. Frescura, F., Giorni, M., Feci, C., Cacopardi, S.: JPEG2000 and MJPEG2000 transmission in 802.11 wireless local area networks. IEEE Trans. Consumer Electron. 11 (2003) 861-871
3. Ghosh, A., Wolter, D. R., Andrews, J. G., Chen, R.: Broadband wireless access with WiMax/802.16: current performance benchmarks and future potential. IEEE Commun. Magazine 2 (2005) 129-136
4. ISO/IEC JTC1/SC29/WG1 WG1N3366: JPEG 2000 image coding system – Part 11: Wireless JPEG 2000 – Committee Draft version 0.1. 7 (2004)
5. Kim, J., Mersereau, R. M., Altunbasak, Y.: Error-resilient image and video transmission over the internet using unequal error protection. IEEE Trans. Image Processing 2 (2003) 121-131
6. ISO/IEC JTC1/SC29/WG1 N2117: Motion JPEG2000 Final Committee Draft 1.0. 3 (2001)
7. Sharma, G., Hassan, A. A., Dholakia, A.: Performance evaluation of burst-error-correcting codes on a Gilbert-Elliot channel. IEEE Trans. On Commun. 7 (1998) 846-849
8. Skodras, A., Christopoulos, C., Ebrahimi, T.: The JPEG 2000 still image compression standard. IEEE Sig. Proc. Mag. 7 (2001) 36- 58
9. Stankovic, V. M., Hamzaoui, R., Charfi, Y., Xiong, Z.: Real-time unequal error protection algorithms for progressive image transmission. IEEE J. Select. Areas Commun. 12 (2003) 1526- 1535
10. Wu, D., Hou, Y. T., Zhu, W., Zhang, Y. Q., Peha, J. M.: Streaming video over the Internet: approaches and directions. IEEE Trans. Circuits Syst. Video Technol. 3 (2001) 282-300

An Ontology Infrastructure for Multimedia Reasoning*

Nikolaos Simou[1], Carsten Saathoff[2], Stamatia Dasiopoulou[3],
Evangelos Spyrou[1], Nikola Voisine[3], Vassilis Tzouvaras[1],
Ioannis Kompatsiaris[3], Yiannis Avrithis[1], and Steffen Staab[2]

[1] Image, Video and Multimedia Systems Laboratory, School of Electrical and
Computer Engineering, National Technical University of Athens,
GR-15773 Zographou Athens, Greece
[2] Informatics and Telematics Institute Centre for Research and Technology Hellas,
GR 57001 Thermi-Thessaloniki, Greece
[3] University of Koblenz, Institute for Computer Science,
D-56016 Koblenz, Germany

Abstract. In this paper, an ontology infrastucture for multimedia reasoning is presented, making it possible to combine low-level visual descriptors with domain specific knowledge and subsequently analyze multimedia content with a generic algorithm that makes use of this knowledge. More specifically, the ontology infrastructure consists of a domain-specific ontology, a visual descriptor ontology (VDO) and an upper ontology. In order to interpret a scene, a set of atom regions is generated by an initial segmentation and their descriptors are extracted. Considering all descriptors in association with the related prototype instances and relations, a genetic algorithm labels the atom regions. Finally, a constraint reasoning engine enables the final region merging and labelling into meaningful objects.

1 Introduction

Recently, there is a growing research interest in the extraction of high-level semantic concepts from images and video using low-level multimedia features and domain knowledge. Significant progress has been made on automatic segmentation or structuring of multimedia content and the extraction of low-level features within such content [1]. However, comparatively little progress has been made on interpretation and generation of semantic descriptions of visual information. More importantly, most analysis techniques focus on specific application domains, making it hard to generalize in case other domains need to handled.

Due to the limitations of the state of the art multimedia analysis systems [2], it is acknowledged that in order to achieve semantic analysis of multimedia content, ontologies [3] are essential to express semantics in a formal machine-processable

* This research was partially supported by the European Commission under contract FP6-001765 aceMedia.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 51–60, 2006.
© Springer-Verlag Berlin Heidelberg 2006

representation. Ontology-based metadata creation currently addresses mainly textual resources or simple annotation of photographs [4]. In well-structured applications (e.g. sports and news broadcasting) domain-specific features that facilitate the modelling of higher level semantics can be extracted [5]. A priori knowledge representation models are also used to assist semantic-based classification and clustering [6]. However, most such techniques are either not suitable for multimedia content analysis, or too correlated with the specific domains they are designed for.

In [7] a novel framework for video content understanding that uses rules constructed from knowledge bases and multimedia ontologies is presented. In [8], multimedia ontologies are semi-automatically constructed using a data-driven approach. [9] presents automatic techniques for extracting semantic concepts and discovering semantic relations among them and evaluates several techniques for visual feature descriptors extraction. In [10], semantic entities, in the context of the MPEG-7 standard, are used for knowledge - assisted video analysis and object detection, while in [11] MPEG-7 compliant low-level descriptors are mapped to intermediate-level descriptors forming an object ontology. It is evident from all such approaches, that a generic multimedia content analysis framework is required that makes use of knowledge stored in multimedia-enabled ontologies.

The framework presented in this paper combines low-level visual descriptors and domain-specific knowledge represented in an ontology infrastructure with a generic analysis scheme to semantically interpret and annotate multimedia content. The infrastructure consists of (i) a domain-specific ontology that provides the necessary conceptualizations for the specific domain, (ii) multimedia ontologies that model the multimedia layer data in terms of low level features and media structure descriptors, and (iii) a core ontology (DOLCE) that bridges the previous ontologies in a single architecture. During image/video analysis, a set of atom-regions is generated by an initial segmentation, and MPEG-7 visual descriptors are extracted for each region. A distance measure between these descriptors and the ones of the prototype instances included in the domain ontology is estimated using a neural network approach. A genetic algorithm then decides the initial labelling of the atom regions with a set of hypotheses, where each hypothesis represents a class from the domain ontology. Finally, a constraint reasoning engine enables the final merging of the regions, while at the same time reducing the number of hypotheses. This approach is generic and applicable to any domain as long as new domain ontologies are designed and made available.

The remainder of the paper is structured as follows: section 2 describes the ontology infrastructure. Section 3 describes the Genetic Algorithm approach, while section 4 describes the proposed reasoning engine. Results are presented in section 5 and conclusions are drawn in section 7.

2 Ontology Infrastructure

There are two main factors that breed the need for a knowledge infrastructure for multimedia analysis. Firstly the fact that reasoners have to deal with large numbers of instantiations of the concepts an properties defined in ontologies, in cases of reasoning with multimedia data on large scale, and secondly that multimedia data comes in two separate though intertwined layers, multimedia and content layer. The multimedia layer deals with the semantics of properties related to the representation of content within the media data itself while on the other hand the content layer deals with the semantics of the actual content contained in the media data as it is perceived by the human media consumer.

Hence the knowledge infrastructure should model the multimedia layer data so as to support extraction and inferencing of content layer data. The ontology infrastructure used integrates these two layers consisting of:

- *Multimedia ontologies* that model the multimedia layer data in terms of low level features and media structure descriptors, namely the *Visual Descriptors Ontology* (VDO), based on an RDF representation of the MPEG-7 Visual Descriptors, and the *Multimedia Structure Ontology* (MSO), based on the MPEG-7 MDS.
- *Domain Ontologies* that provide the necessary conceptualizations of the content layer, for a specific application domain.
- A *Core Ontology* that models primitives at the root of the concept hierarchy and can be exploited by both types of ontologies. It is also meant to bridge between the other ontologies within the architecture.

The knowledge infrastructure is set up using RDFS. This approach is expected to be complemented by using an appropriate sub-language of OWL at a later stage. This decision reflects that a full usage of the increased expressiveness of OWL requires specialized and more advanced inference engines, especially when dealing with large numbers of instances.

The Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) was explicitly designed as core ontology. The RDFS version of DOLCE currently contains about 79 high level concepts and 81 high level properties among them. DOLCE contains explicit conceptualizations by including the concept of qualities that can be perceived, as well as spatio-temporal concept descriptions. However, reasoning with spatio-temporal descriptions requires the coding of additional relations that describe the relationship between space and/or time regions. Based on concepts taken from Region Connecting Calculus, Allen's interval calculus and directional models, we have carefully extended DOLCE to accomodate the corresponding directional and topological relationships in the spatial and temporal domains.

The top-level multimedia content entities of the MSO are described in MPEG-7 Multimedia Description Schemes (MDS) FCD. Within MPEG-7, multimedia content is classified into five types: Image, Video, Audio, Audiovisual and Multimedia. Each of these types has its own segment subclasses. The Segment DS describes a spatial and/or temporal fragment of multimedia content. A number of specialized subclasses are derived from the generic Segment DS. These subclasses describe the specific types of multimedia segments, such as video segments, moving regions, still regions and mosaics, which result from spatial, temporal and spatiotemporal segmentation of the different multimedia content types. Multimedia resources can then be accordingly decomposed into sub-segments through spatial, temporal, spatiotemporal or media source decomposition.

The VDO contains a set of visual descriptors to be used for knowledge-assisted analysis of multimedia content. By the term descriptor we mean a specific representation of a visual feature (color, shape, texture etc) that defines the syntax and the semantics of a specific aspect of the feature (dominant color, region shape etc). The entire VDO follows closely the specification of the MPEG-7 Visual Part, but several modifications were carried out in order to adapt to the datatype representations available in RDFS.

In order to extract a set of prototype low-level visual descriptors for different domain concepts and integrate them into the ontology structure, it must be clear how domain concepts can be linked with actual instance data without having to cope with meta-modelling. For this purpose, we have enriched the knowledge base with instances of domain concepts that serve as *prototypes* for these concepts. Each of these is linked to the appropriate visual descriptor instances.

3 Knowledge-Assisted Analysis

The domain ontology represents the required knowledge for interpreting each image or video scene, which is a mapping of image regions to the corresponding domain-specific semantic definition. Classes within the ontology have been defined to represent the different types of visual information while subclasses represent the different ways to calculate a visual feature. Each real-world object is allowed to have more than one instantiations. Currently, three spatial relations and three low-level descriptors are supported. These descriptors are: adjacency (ADJ), below (BEW), and inclusion (INC) relations, and dominant color (DC), motion (MOV) and compactness (CPS) descriptors. Enriching the ontology with domain specific knowledge results in populating it with appropriate instances, i.e. prototypes for the objects to be detected.

During preprocessing, color segmentation [12][1]) and motion segmentation [13][11]) are combined to generate a set of over-segmented atom-regions. The extraction of the low-level descriptors for each atom-region is performed using the MPEG-7 eXperimentation Model(XM) [1]. Motion estimation is based on block motion vector estimation using block matching and the calculation of the norm of the averaged global-motion-compensated motion vectors for the blocks belonging to each region. Global motion compensation is based on estimating the 8 parameters of the bilinear motion model for camera motion, using an iterative rejection procedure [14]. Finally, the compactness descriptor is calculated by the area and the perimeter of the region.

After preprocessing, assuming for a single image N_R atom regions and a domain ontology of N_O objects, there are $N_R^{N_O}$ possible scene interpretations. A genetic algorithm is used to overcome the computational time constraints of testing all possible configurations [15]. In this approach, each individual represents a possible interpretation of the examined scene, i.e the identification of all atom regions. In order to reduce the search space, the initial population is generated by allowing each gene to associate the corresponding atom-region only with those objects that the particular atom-region is most likely to represent.

The degree of matching between regions, in terms of low-level visual and spatial features respectively, is defined as:

- the interpretation function $\mathcal{I}_M(g_i) \equiv \mathcal{I}_M(R_i, om_j)$, assuming that g_i associates region R_i with object o_j having model om_j, to provide an estimation of the degree of matching between an object model om_j and a region R_i. $\mathcal{I}_M(R_i, om_j)$ is calculated using the descriptor distance functions realized in the MPEG-7 XM and is subsequently normalized so that $\mathcal{I}_M(R_i, om_j)$ belongs to $[0, 1]$.
- the interpretation function $\mathcal{I}_{\mathcal{R}}$, which provides an estimation of the degree to which a relation $\mathcal{R}$ holds between two atom-regions.

The employed fitness function that considers the above matching estimations for all atom-regions is defined as:

$$Fitness(G) = \sum_{g_i} \mathcal{I}_M(g_i) + \sum_{k} \sum_{(g_i, g_j)} \mathcal{I}_{\mathcal{R}_k}(g_i, g_j)$$

where $\mathcal{I}_M(g_i)$ is the estimation function of gene g_i regarding low-level visual similarity and $\mathcal{I}_{\mathcal{R}_k}(g_i, g_j)$ is the estimation function of spatial similarity between g_i and g_j in terms of $\mathcal{R}_k$. It follows from the above definitions that the optimal solution is the one that maximizes the fitness function. Any neighboring regions belonging to the same object according to the generated optimal solution are simply merged. For each object that fails to comply the concept of unknown object is introduced.

Our approach to implement the interpretation function $\mathcal{I}_M$ used for the fitness function is based on a back - propagation neural network. When the task is to compare two regions based on a single descriptor, several distance functions can be used; however, there is not a single one to include all descriptors with different weight on each. This is a problem that the neural network handles. Its input consists of the low-level descriptions of both of an atom region and an object model, while its response is the estimated normalized distance between the atom region and the model. A training set is constructed using the descriptors of a set of manually labelled atom regions and the descriptors of the corresponding object models. The network is trained under the assumption that the distance of an atom region that belongs to the training set is minimum for the associated object and maximum for all others. This distance is then used for the interpretation function $\mathcal{I}_M$.

4 Constraint Reasoning Engine

The analysis procedure described in section 3 results in an image segmented into a number of atom regions, each labeled with an initial set of hypotheses. Each hypothesis corresponds to one object description defined in the domain ontology. Although at this stage the atom-regions bear semantic information, further processing is required to derive a segmentation where each segment represents a meaningful object. To accomplish this, the limitations posed by the numerically based segmentation algorithms need to be overcome, i.e. atom-regions corresponding to only part instead of the complete object, loss of object connectivity etc. In the following we describe an approach to meet this requirements based on reasoning on the labels and spatiotemporal information of the considered atom-regions.

The input of the proposed reasoning system consists of the set of atom-regions along with their initial labels as resulted following the former analysis procedure. The corresponding output is a reduced number of atom-regions, which coincide with real objects, within a plausible degree of accuracy, and a reduced set of hypotheses for each atom-region. The reasoning process is based on the extracted labels and spatiotemporal features of the examined atom-regions in association with the information included in the domain ontology.

The integration of low-level features in the reasoning process further improves the plausibility of the detection results. For example, a merging indicated by the defined rules should be performed only if the shape of the resulting segment conforms to the shape of one of the plausible object classes corresponding to the merged atom-regions. Obviously, this raises the need for incorporation of low-level feature matching into the reasoning system, which on the one hand can lead to computational problems and on the other hand reduces the number of eligible reasoning systems, because means to extend the system must be available.

This can be better understood through the example of Fig. 1, which is initially segmented and labelled as illustrated in Fig. 2. Regions labelled as ‘sky’, ‘field’ and ‘mountain’ are expected to be merged. Furthermore, more complex regions such as ‘roof’ and ‘wall’ are evaluated in association with each other. In other words, since a region has bright red color and geometrically fits to the

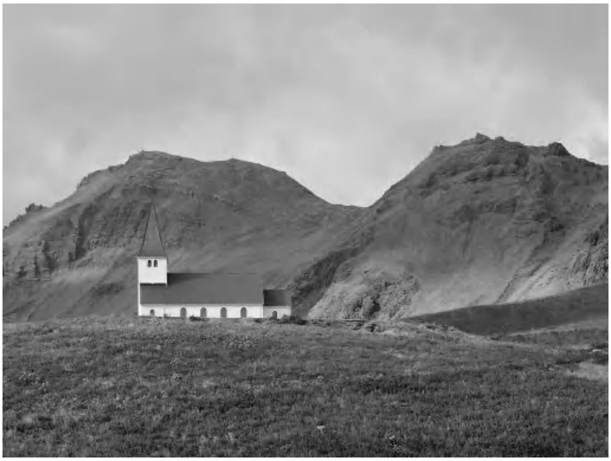


Fig. 1. Input image

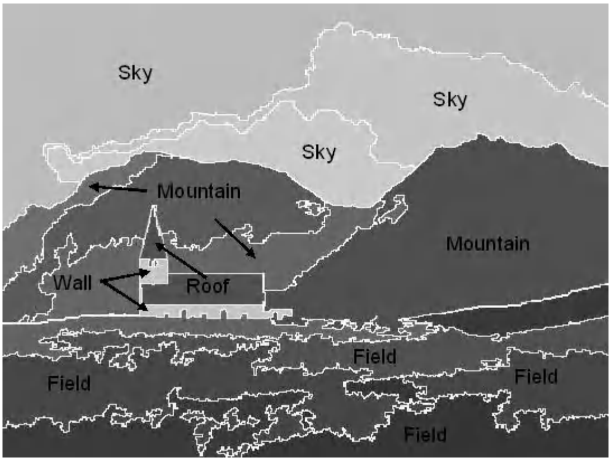


Fig. 2. Image after initial segmentation and labelling

description given for a roof it may be labelled as ‘roof’. On the other hand, a white rectangle is difficult to be assigned a label alone, due to its very general features appearing in a number of prototype instances; however, according to available spatiotemporal information (white rectangle below roof) the ‘wall’ label is assigned to it.

The whole process is iterative, as the actual region merging cannot be implemented efficiently within a reasoning system. Thus, the reasoner identifies regions that are to be merged, by adding a relation between them. Such relations are interpreted within a second step, where regions are merged, and any associated visual descriptors and relations are updated. New region labels are

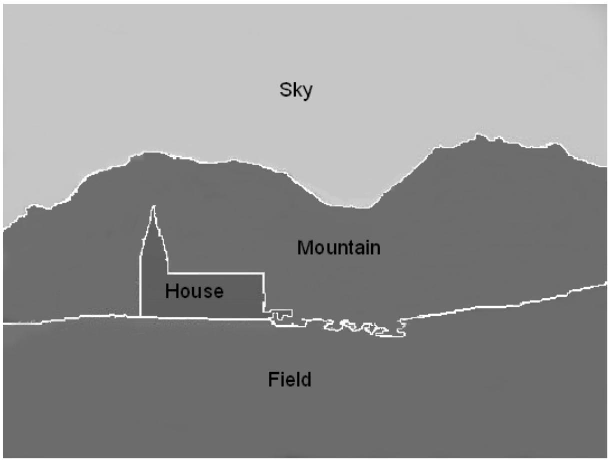


Fig. 3. Output of the constraint reasoner

estimated, and hypotheses are then constructed. The output of this analysis step again serves as input for the reasoner in an iterative fashion until a stable state is reached, i.e. no new information can be inferred.

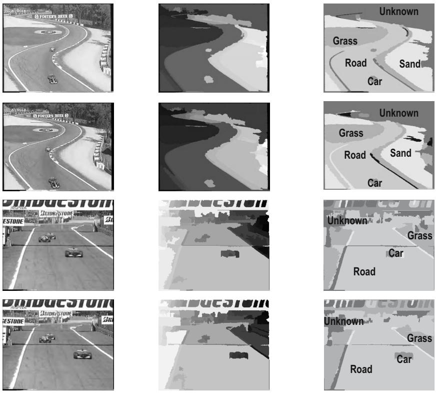
Using this approach, objects with similar visual characteristics can be discriminated in terms of their spatiotemporal behavior and the visual context on which they occur. Furthermore, based on the output of the described reasoning process, further analysis becomes feasible, aiming at the generation of higher-level semantics, such as recognition of complex objects or events, which cannot be represented in terms of their visual features. In our example this is shown at Fig. 3. The 'sky', 'field' and 'mountain' regions have been merged but also regions 'wall' and 'roof' have been merged with the label 'house' assigned to the resulting region.

5 Results

The presented ontology-based framework was used to extract semantic descriptions of a variety of MPEG-2 videos of the Formula One and Tennis domains. The corresponding domain ontologies, i.e the defined object classes along with

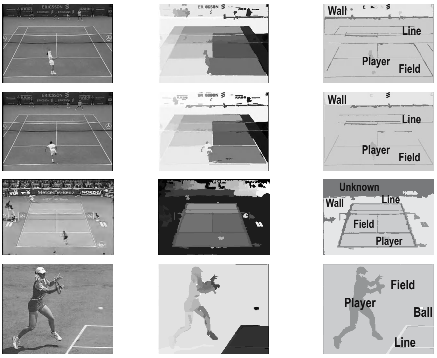
Table 1. Formula One and Tennis domain definitions

Object Class	Low-level descriptors	Spatial relations
Road	$DC^1_{road} \vee DC^2_{road} \vee DC^3_{road}$	Road ADJ Grass,Sand
Car	$MOV^1_{car} \wedge CPS^1_{car}$	Car INC Road
Sand	$DC^1_{sand} \vee DC^2_{sand}$	Sand ADJ Grass, Road
Grass	$DC^1_{grass} \vee DC^2_{grass} \vee DC^3_{grass}$	Grass ADJ Road,Sand
Field	$DC^1_{field} \vee DC^2_{field} \vee DC^3_{field}$	Field ADJ Wall
Player	MOV^1_{Player}	Player INC Field
Line	$DC^1_{line} \wedge CPS^1_{line}$	Line INC Field
Ball	$DC^1_{Ball} \wedge CPS^1_{Ball}$	Ball INC Field
Wall	$DC^1_{Wall} \vee DC^2_{Wall} \vee DC^2_{Wall}$	Wall ADJ Field



Input Images Segmentations Interpretations

Fig. 4. Formula One domain results



Input Images Segmentations Interpretations

Fig. 5. Tennis domain results

their low-level features and spatial interrelations are illustrated in Table 1. A training set of manually annotated videos was used to populate the domain ontologies with prototype instances.

As illustrated in Fig. 4 and 5, the system output is a segmentation mask outlining the semantic description of the scene where different colors representing the object classes defined in the domain ontology are assigned to the generated atom-regions.

Excluding the process of motion information extraction, the required analysis time was between 5 and 10 seconds per frame. The use of spatial information captures part of the visual context, consequently resulting in the extraction of more meaningful descriptions provided that the initial color-based segmentation did not segment two objects as one atom-region.

6 Conclusion

In this paper an approach is described that combines multimedia domain knowledge, a knowledge-assisted analysis and a constraint reasoner in order to extract high-level semantic knowledge from images and video. The developed ontology infrastructure efficiently relates domain knowledge with the semantics of the visual part of MPEG-7 through an upper harmonizing ontology. After a preprocessing step, a genetic algorithm generates a set of region label hypotheses, which are then processed by a constraint reasoning engine in an iterative fashion to enable the final region merging and labelling. The entire approach is generic, in the sense that all domain-specific information solely resides in the domain ontology; the same analysis framework has been tested on two different domains simply by switching the associated domain ontology, with promising initial results. This research is ongoing and future work includes implementation of larger scale domain ontologies enhanced with multimedia descriptions, relations and rules to evaluate the proposed methodology on a large set of multimedia content.

References

1. Manjunath, B., Ohm, J.R., Vasudevan, V., Yamada, A.: Color and texture descriptors. IEEE Trans. on Circuits and Systems for Video Technology, special issue on MPEG-7 **11**(6) (2001) 703–715
2. R. Brunelli, O.M., Modena, C.: A survey on video indexing. Journal of Visual Communications and Image Representation **10** (1999) 78–112
3. Staab, S., Studer, R.: Handbook on Ontologies. International Handbooks on Information Systems. Springer Verlag, Heidelberg (2004)
4. A.Th. Schreiber, B. Dubbeldam, J.W., Wielinga, B.: Ontology-based photo annotation. IEEE Intelligent Systems (2001)
5. Al-Khatib, W., Day, Y., Ghafoor, A., Berra, P.: Semantic modeling and knowledge representation in multimedia databases. IEEE Transactions on Knowledge and Data Engineering **11**(1) (1999) 64–80
6. Yoshitaka, A., Kishida, S., Hirakawa, M., Ichikawa, T.: Knowledge-assisted content-based retrieval for multimedia databases. IEEE Multimedia **1**(4) (1994) 12–21
7. Alejandro Jaimes, B.T., Smith, J.R. (In: Proc. IEEE International Conference on Image and Video Retrieval (ICME 2003))
8. Jaimes, A., Smith, J.R. (In: Proc. IEEE International Conference on Multimedia and Expo (ICME 2003))
9. Benitez, A.B., Chang, S.F. (In: Proc. IEEE International Conference on Image and Video Retrieval (ICME 2002))
10. Tsechpenakis, G., Akrivas, G., Andreou, G., Stamou, G., Kollias, S.: Knowledge-Assisted Video Analysis and Object Detection. In: Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (Eunite02), Algarve, Portugal (2002)
11. Mezaris, V., Kompatsiaris, I., Boulgouris, N., Strintzis, M.: Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. IEEE Trans. on Circuits and Systems for Video Technology **14**(5) (2004) 606–621
12. Mezaris, V., Kompatsiaris, I., Strintzis, M.: A framework for the efficient segmentation of large-format color images. In: Proc. International Conference on Image Processing. Volume 1. (2002) 761–764
13. Tuan, J.C., Chang, T.S., Jen, C.W.: On the data reuse and memory bandwidth analysis for full-search block-matching VLSI architecture. IEEE Trans. on Circuits and Systems for Video Technology **12**(1) (2002) 61–72
14. Yu, T., Zhang, Y.: Retrieval of video clips using global motion information. Electronics Letters **37**(14) (2001) 893–895
15. Mitchell, M.: An introduction to Genetic Algorithms. MIT Press. (1996)

Reduced Frame Quantization in Video Coding

Tuukka Toivonen and Janne Heikkilä

Machine Vision Group,
Infotech Oulu and Department of Electrical and Information Engineering,
P.O. Box 4500, FIN-90014 University of Oulu, Finland
{tuukkat, jth}@ee.oulu.fi

Abstract. We vary the quantization parameter in H.264 video encoding by increasing it by a well-chosen offset in every other frame, which we call reduced frames. As the motion compensation for a reduced frame is performed mainly from the previous frame (which is non-reduced), we obtain a large reduction in bit rate with only a small loss in video quality. We then develop an algorithm for the offset selection which requires only slightly more computation compared to the original encoder.

Since the frame immediately preceding a non-reduced frame is reduced, advantage can be obtained by using at least two previous frames for reference. Even larger advantage can be obtained by reordering the reference picture list so that the frame, which gives the best motion compensation result, is first in the list. We also implement an algorithm for deciding the best list order.

The modified encoder produces 6% smaller bit rate in average at fixed video quality at low bit rates, or up to 18% with some sequences which contain relatively little motion.

1 Introduction

Most video coding standards, including MPEG-1/2/4, H.261, H.263, and H.264/AVC [1, 2], require hybrid encoders which first perform motion compensation by predicting picture element (pixel) values from previously encoded frames, followed by a residual transform using a discrete cosine transform (DCT) or its approximation. The transform coefficients are quantized by dividing them by a quantization parameter Q_0 and rounding them to integers. The integers are then sent to a decoder together with motion compensation information (motion vectors).

The objective of video coding is to obtain optimal rate-distortion behavior, that is, given a bit rate, the quality should be as high as possible, or equivalently, for a given quality the bit rate should be as low as possible. In standards-based encoders the tradeoff between bit rate and quality is controlled with Q_0. By increasing it, the coefficients are quantized more strongly, and more of them will be zero, requiring less bits. However, the quality of the video sequence is decreased.

If a frame can be accurately predicted from a previous frame with motion compensation, the transform coefficients could be quantized very coarsely, with

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 61–67, 2006.
© Springer-Verlag Berlin Heidelberg 2006

the result still being good. This approach is often used with bidirectionally predicted B-frames by increasing Q_0 slightly [3]. Except that being accurately predicted, there is another important reason why B-frames can be quantized coarsely without affecting much the overall quality: B-frames are not usually used in motion compensation of later frames, and therefore low quality B-frames will not decrease the quality of subsequent frames. This suggests that more common unidirectionally predicted P-frames soon after an intra-coded I-frame should be encoded with higher fidelity than P-frames just before an I-frame [4], to prevent propagation of large quantization noise to many frames.

Our method utilizes the same idea by adding a well-chosen constant ΔQ to Q_0 in every other P-frame. We call these frames as reduced frames. Because the reduced frames are motion compensated mainly from the immediately preceding frame, which was encoded at full quality, the reduced frames will have only a small quality loss. In the H.264 standard, the reference frame for motion compensation can be selected. Therefore, motion compensation of non-reduced frames can be done using the previous high quality frame, avoiding the quantization error propagation from a reduced frame.

In this paper we develop an algorithm for automatic selection of the quantization parameter offset ΔQ and apply it to P-frames, thus avoiding the use of B-frames and maintaining compatibility with H.264 baseline profile. While we obtain the best advantage by using at least two previous frames for motion compensation, some improvement is achieved even with just one reference frame, making the algorithm and the technique in general useful also with older video coding standards.

For H.264, we also implement an algorithm for reordering the reference picture list so that the frame, which usually gives the best motion compensation result, is first in the list. This allows using shorter codes for frame reference in the encoded bit stream, further decreasing the bit rate.

2 Algorithm

Our algorithm for the selection of ΔQ encodes reduced frames several times. First, a frame is encoded without applying the offset ($\Delta Q = 0$) with quantization parameter Q_0. As a result, the distortion D_0 and the bit rate R_0 corresponding to the case of zero offset can be calculated. In our work, we measure distortion with mean squared error (MSE) and the bit rate using average number of bits per pixel (BPP).

Then, the frame is encoded again using the best offset ΔQ_b determined previously. Initially we choose $\Delta Q_b = 1$. From this second pass, we get D_b and R_b corresponding to quantization parameter $Q_b = Q_0 + \Delta Q_b$.

We need to know whether the offset improved the encoding result, and if it did, by how much. This can be accomplished by computing $\widehat{D}_b$, corresponding to D_b but at the bit rate R_0 instead of R_b, as

$$\widehat{D}_b = (R_0 - R_b)\,\kappa + D_b \tag{1}$$

where κ is the rate-distortion curve slope at the current operational point, depending on Q_0. We describe later how it can be estimated. With this result, we compute the distortion difference of the two first passes at the same bit rate as

$$\Delta D_b = \widehat{D}_b - D_0. \tag{2}$$

The frame is encoded third time with the offset $\Delta Q_{b+1} = \Delta Q_b + 1$ increased by one which produces slightly higher distortion D_{b+1} and lower bit rate R_{b+1}. The distortion and the distortion difference are computed similarly as before: $\widehat{D}_{b+1} = (R_0 - R_{b+1})\,\kappa + D_b$ and $\Delta D_{b+1} = \widehat{D}_{b+1} - D_0$.

After calculating ΔD_b and ΔD_{b+1}, we know if the rate-distortion behavior improved $(\Delta D_b > \Delta D_{b+1})$ or worsened $(\Delta D_b < \Delta D_{b+1})$ by increasing the offset. By assuming that the distortion $\Delta D_{b+\Delta}$ is unimodal, we continue searching the minimum by increasing or decreasing the offset until the distortion gets worse (larger) instead of better. Then, a step is taken backward to the local optimum and the frame is encoded with the final offset value. This quantization parameter corresponds to the point on the rate-distortion curve of the frame where its slope is κ.

The implemented algorithm modifies Q_0 only as used in quantization. The Lagrangian multiplier λ which is used in motion estimation and macroblock mode selection [5] is derived from the original quantization parameter before the offset is added to it. Otherwise the higher quantization would decrease also motion compensation quality by favoring larger partition sizes and less accurate motion vectors.

The parameter κ approximates the rate-distortion curve slope. We derived it by encoding 13 different video sequences with nine different fixed Q_0 values between 20 and 36 (see Section 3), without applying any offset. The rate-distortion curve for each sequence was then approximated by interpolating a cubic spline through the data points. The spline derivative was evaluated to obtain the curve slope at each data point. Finally a least squares estimation was used to find the optimal parameters in

$$\kappa = -2^{\alpha Q_0 + \beta} \tag{3}$$

where κ is the curve slope at quantization parameter Q_0. As a result, we obtained $\alpha = 0.4315$ and $\beta = -5.704$, as shown in Fig. 1. The parameter κ is closely related to the Lagrangian multiplier: in theory, $\lambda = -\kappa$. In our implementation, however, we did not modify the original selection of λ, which is usually computed as $\lambda = 0.85 \times 2^{(Q_0-12)/3} \approx 2^{0.33 \times Q_0 - 4.23}$.

As mentioned in Section 1, the encoding results can be improved significantly if at least two previous frames are available for motion compensation so that a non-reduced frame can predict from the previous non-reduced frame and if the reference picture list is reordered so that the best reference frame is first in the list. With a list containing two frames we have only two alternatives: either the pictures are swapped in the list, bringing the previous non-reduced frame first, or they are left in the encoding order. We perform the decision using the same principle as we search for the optimum quantization parameter offset: by encoding a frame twice, with the two entries in the list swapped and not swapped. The alternative giving smaller distortion at the original rate is then chosen.

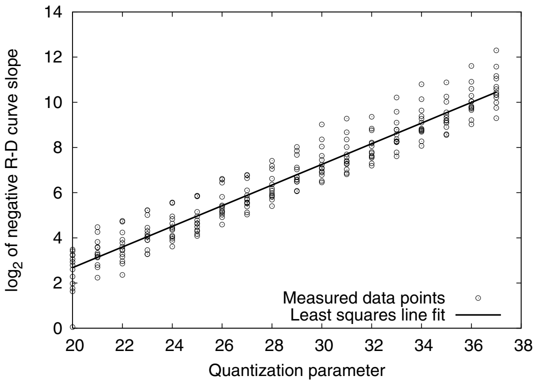


Fig. 1. Rate-distortion curve slope at various quantization parameter values and a least squares fit

We did not optimize the algorithms for speed, and if optimum offset is searched as described above, each reduced frame has to be encoded at least four times, requiring four times longer encoding time. However, there is a simple way to improve the execution time to very near the original encoder: according to our experiments (see Section 3), the optimum offset and frame order typically do not fluctuate much. Thus, they can be searched only rarely, for example after every 250 frames, while still obtaining similar encoding results but with only slightly more computation than in the original encoder.

3 Experimental Results

The algorithms were implemented into H.264 encoder x264[1] revision 239 and tested with 12 CIF-sized (352×288) standard test sequences, listed in Table 1, each 250 frames long. Most H.264 quality-improving features were enabled, such as CABAC and loop filter. In the tests we used either one or two reference frames, and the sequence format was IPPP. Each sequence was encoded with nine different fixed quantization values between 20 and 36. Fig. 2 displays the frame bit rate and PSNR for 50 consecutive frames of *Paris* sequence with $Q_0 = 28$ and ΔQ_b computed adaptively. As can be seen, although the bit rate is reduced enormously, the quality is decreased only slightly in reduced frames.

Figures 3a and b display the relative bit rate reduction at constant PSNR level compared to the original encoder (at 100%, case 1), for the optimum offset searched for every reduced frame (case 2), and for the optimum offset searched only once in the beginning of each sequence (case 4), averaged over all 12 sequences. Fig. 3a displays the results obtained with two frames in the reference picture list, while in Fig. 3b test cases only one reference frame is made available,

[1] Available from http://www.videolan.org/x264.html. Although x264 is very fast (real time), the coding quality is comparable to the H.264 test model JM 9.6. See http://www.ee.oulu.fi/˜tuukkat/mplayer/tests/readme.html.

Table 1. Test sequences at $Q_0 = 28$

Sequence	Offset mean	Swapped
Bridge Close	3.63	8.1%
Coastguard	3.46	50.4%
Flower Garden	3.57	68.5%
Highway	4.16	40.7%
News	4.35	20.2%
Foreman	4.33	59.3%
Mobile and Calendar	4.13	99.2%
Munchener Hall	4.03	43.1%
Paris	4.03	21.1%
Stefan	4.31	63.4%
Tempete	4.22	97.6%
Waterfall	6.07	99.2%
Average	4.19	55.9%

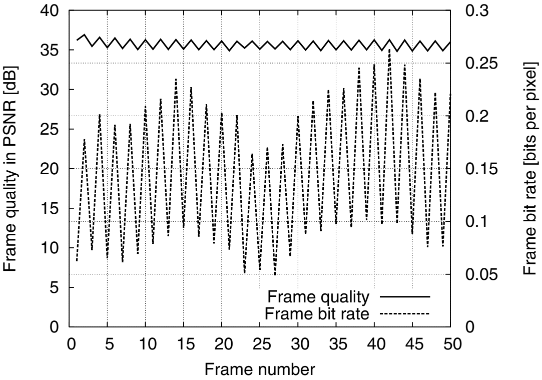
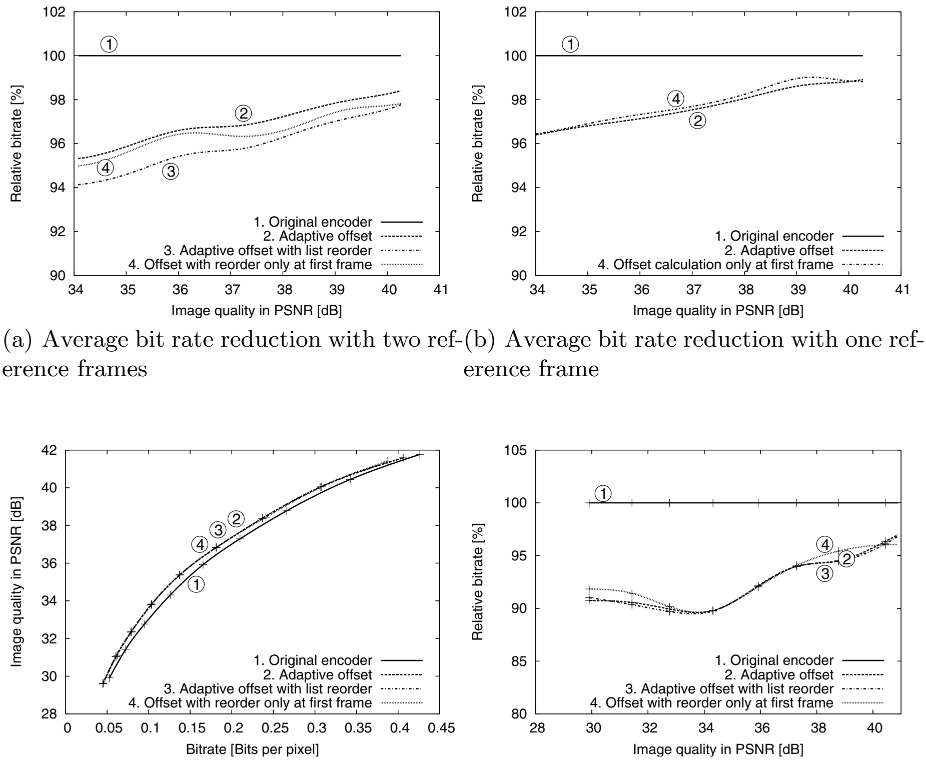


Fig. 2. Frame PSNR and bit rate versus frame number in *Paris* at $Q_0 = 28$

similarly as in older standards. Fig. 3a displays also the improvement, which is obtained by reordering the reference picture list into the best order (case 3). In the last test case 4, also the optimum order is searched only once in the beginning of each sequence and the same order is used for the rest of the frames. Fig. 3c displays the rate-distortion curves in the four cases for one particular test sequence, *Paris*, with two reference frames, and Fig. 3d displays the corresponding relative bit rates, computed from results shown in Fig. 3c.

As can be seen in the figures, the method gives best results at low bit rates corresponding to low PSNR levels, where up to 6% decrease in bit rate is achieved with two reference frames, or up to 3.4% decrease if only one reference frame is available. However, the results vary much depending on sequence. With the *Waterfall* sequence, even 18% reduction in bit rate is achieved.

By computing the optimum offset and whether the two reference frames are swapped only once during a whole video sequence, computation demand is only about 3% higher than in the original encoder while bit rate is reduced up to 5% and 3.4% with two or one reference frames, respectively.



(a) Average bit rate reduction with two reference frames

(b) Average bit rate reduction with one reference frame

(c) Rate-distortion curve of *Paris*

(d) Relative bit rate of *Paris*

Fig. 3. Relative bit rate reduction in average and for one particular sequence with rate-distortion curve

Table 1 displays the average quantization parameter offset for each test sequence at $Q_0 = 28$, when the optimization algorithm was ran for every frame. It also shows in how large percentage of non-reduced encoded frames the two entries in the reference list were swapped.

4 Conclusions

We observed a significant improvement in rate-distortion behavior when a well-chosen offset is added to the quantization parameter Q_0 at every other P-frame. This can be explained by noticing that while a larger Q_0 decreases significantly encoded frame size, motion compensation can still produce a frame which is near the original as it uses a higher quality frame for prediction.

Based on this observation, we developed an algorithm for selecting the offset automatically. We also considered arranging reference picture list so that the high quality frames giving the best prediction are first in the list. The total bit rate is reduced in average by 6% at low bit rates, and up to 18% with some video sequences which contain only little motion. The additional computation cost is quite high, as each frame must be encoded multiple times. However, with simple modification the computation is decreased so that only around 3% more is needed compared to the original encoder while still obtaining 5% average reduction in bit rate.

The method works best when at least two previously encoded frames are used for motion compensation which is generally possible only with the newest standard H.264. However, some improvement is achieved also with just one reference frame, making the method usable with older standards as well.

References

1. Joint Video Team of ITU-T and ISO/IEC JTC 1: ITU-T Recommendation and International Standard of Joint Video Specification. ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC (2005).
2. Wiegand, T., Sullivan, G. J., Bjontegaard, G., Luthra,A.: Overview of the H.264/AVC Video Coding Standard. IEEE Trans. Circuits Syst. Video Technol **13**, no. 7 (2003) 560–576.
3. Flierl, M., Girod, B.: Generalized B Pictures and the Draft H.264/AVC Video Compression Standard. IEEE Trans. Circuits Syst. Video Technol **13**, no. 7 (2003) 587–597 .
4. Feng, P., Li,Z., Lim, K., Feng, G.: A Study of MPEG-4 Rate Control Scheme and Its Improvements. IEEE Trans. Circuits Syst. Video Technol **13**, no. 5 (2003) 440–446.
5. Sullivan, G. J., Wiegand, T.: Rate-Distortion Optimization for Video Compression. IEEE Signal Processing Mag. **15**, no. 6 (1998) 74–90.

Video Object Watermarking Based on Moments

Paraskevi K. Tzouveli, Klimis S. Ntalianis, and Stefanos D. Kollias

National Technical University of Athens,
Electrical and Computer Engineering School, Athens, Greece
tpar@image.ntua.gr

Abstract. A robust video object based watermarking scheme, based on Zernike and Hu moments, is proposed in this paper. Firstly, a human video object detector is applied to the initial image. Zernike and the Hu moments of each human video object are estimated and an invariant function for watermarking is incorporated. Then, the watermark is generated modifying the moment values of each human video object. In the detection scheme, a neural network classifier is initially used in order to extract possible watermarked human video objects from each received input image. Then, a watermark detection procedure is applied for video object authentication. A full experiment confirms the promising performance of the proposed scheme. Furthermore, the performances of the two types of moments are extensively investigated under several attacks, verifying the robustness of Zernike moments comparing to Hu moments.

1 Introduction

Information security is one of the most significant and challenging problems that modern societies confront. This is due to the fact that anyone can easily copy digital data, reproduce it without information loss, and manipulate it without detection. Whether or not a multimedia system is sufficiently secure will have a substantial influence on its acceptance and spreading throughout communication networks and applications. For example, security solutions that address the fields of distributed production and e-commerce are especially necessary because they provide access control mechanisms to prevent misuse and theft. On the other hand, the intellectual property protection problem is also of crucial importance. Several watermarking methods have been presented in literature, some of which are based on moments [1] - [4]. Moments and invariant functions of moments have been investigated for invariant feature extraction and modification of these features.

In [4] from the regular moments, Hu derived seven moment functions which are rotation, scaling and translation invariant. In [7], Hu's moment invariant functions are used for watermarking. The watermark is embedded by modifying the moment values of the image. In their implementation, an exhaustive search should be performed towards determination of the embedding strength. Teague, [6], proposed Zernike moments based on the basis set of orthogonal Zernike polynomials [5]. Zernike moments are powerful feature descriptors which can easily provide rotation invariance. The utility of Zernike moments in many applications can be considerably enhanced by adding translation and scaling invariance.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 68 – 75, 2006.
© Springer-Verlag Berlin Heidelberg 2006

In this paper, a human video objects watermarking system is designed and implemented. At the first step, for each initial image, automatic human video object detection is performed, based on the method described in [8]. Afterwards, the Hu and Zernike moments of each human video object are estimated and an invariant function is incorporated for watermarking. Embedding of the watermark is achieved by modifying the moment values of each human video object. In the detection scheme, a neural network classifier is initially used in order to extract possible watermarked human video objects from each received input image. Then, a watermark detection procedure is applied for video object authentication. Several experiments are performed which confirm the promising performance of the proposed scheme. Furthermore, the performance of the two types of moments is extensively investigated under several attacks. The paper is organized as follows: in Section II the two types of moments are defined and their properties are briefly summarized. Moreover emphasis is given in Hu and Zernike moments, which are used in the proposed watermarking system. In section III the watermark embedding and detection procedures are analysed while experimental results are presented in section IV.

2 Watermarking and Moments

Geometric moments (regular moments) are non-negative integers, which can be computed by:

$$m_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_p y_q f(x,y)dxdy \tag{1}$$

where m_{pq} is the $(p+q)$th order moment of any real continuous image function $f(x,y)$ having a bounded support and a finite nonzero integral. In a digital implementation, these integrals are approximated by:

$$m_{pq} = \sum_x \sum_y x_p y_q f(x,y) \ . \tag{2}$$

Corresponding *central moment* $\mu_{pq}^{(f)}$ of order $(p+q)$ of the image $f(x,y)$ are defined analogously as

$$\mu_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\bar{x})^p (y-\bar{y})^q f(x,y)dxdy \tag{3}$$

where the coordinates $\bar{x} = \frac{m_{1,0}}{m_{0,0}}, \bar{y} = \frac{m_{0,1}}{m_{0,0}}$ denote the centroids of $f(x,y)$.

The central moments of the image are invariant to translation as they are origin-independent. Scaling invariance can be achieved by normalizing the moments of the scaled image by the scaled energy of the original. For this propose, further normalization for the effects of scaling can be computed by the definition of the formula: $\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}{}^{\gamma}}$ where γ is the normalization factor $\gamma = \frac{p+q}{2} + 1$.

Traditionally, moment invariants are computed based on information provided by both the shape boundary and its interior region [5]-[7].

Zernike, [5], introduced a set of complex polynomials which form a complete orthogonal set over the interior of the unit circle ($x^2+y^2 \leq 1$). The Zernike moments of order *n* (nonnegative integer) with repetition *m* (integer, $n-|m|$ even, $|m| \leq n$) for a continuous image function $f(x,y)$ that vanishes outside the unit circle are

$$A_{nm} = \frac{n+1}{\pi} \int \int_{x^2+y^2 \leq 1} f(x,y) \cdot V_{nm}^{*}(\rho, \vartheta) dxdy \tag{4}$$

Actually, Zernike moments are projections of the image intensity onto the orthogonal basis functions $V_{nm}(x,y) = V_{nm}(\rho, \vartheta) = Rnm(\rho)\exp(jm\theta)$ where ρ and ϑ represent polar coordinates over the unit disk and R_{nm} are the orthogonal polynomials of ρ (Zernike polynomials) given by

$$R_{nm}(\rho) = \sum_{s=0}^{n-|m|/2} \frac{(-1)^s[(n-s)!]\rho^{n-2s}}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} \tag{5}$$

For a digital image, the integrals are replaced by summations. The discrete function $\hat{f}(x,y)$ has exactly the same moments with the $f(x,y)$ up to the given order n_{max}. Zernike moments are the coefficients of the image expansion into orthogonal Zernike polynomials.

Zernike moments acquire a phase shift on rotation and the magnitude $|A'_{nm}|$ of the Zernike moments can be used as a rotation-invariance feature of the image. Scale and translation invariance can be achieved by utilizing the image normalization technique as shown in [1]. If we compute the Zernike moments of the image, then the magnitudes of the moments are RST invariant.

On the other hand, Hu [4] first introduced the mathematical foundation for two-dimensional moment invariants in 1961, based on methods of algebraic invariants and demonstrated their applications to shape recognition [4]. Using nonlinear combinations of geometric moments, a set of invariant moments, which have the desirable properties of being invariant under image translation, scaling, and rotation, is provided by the Hu method.

Hu defines seven of these region descriptor values, computed from central moments through order three, that are independent to object translation, scale and orientation. Translation invariance is achieved by computing moments that are normalised with respect to the centre of gravity so that the centre of mass of the distribution is at the origin (central moments). Size invariant moments are derived from algebraic invariants but these can be shown to be the result of simple size normalization. From the second and third order values of the normalized central moments, a set of seven invariant moments can be computed which are rotation- independent. From the normalized central moments, a set of seven values can be calculated by the following equations:

$$\begin{aligned}
\phi_1 &= n_{20} + n_{02} \\
\phi_2 &= (n_{20} - n_{20})^2 + 4n_{11}^2 \\
\phi_3 &= (n_{30} - 3n_{12})^2 + (n_{03} - 3n_{21})^2 \\
\phi_4 &= (n_{30} - n_{12})^2 + (n_{03} + n_{21})^2 \\
\phi_5 &= (3n_{30} - 3n_{12})(n_{30} + n_{12}) \cdot \left[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2\right] \\
&\quad + (3n_{21} - n_{03})(n_{21} + n_{03}) \cdot \left[3(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2\right] \\
\phi_6 &= (n_{20} - n_{02}) \cdot \left[(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2\right] \\
&\quad + 4n_{11}(n_{30} + n_{12})(n_{21} + n_{03}) \\
\phi_7 &= (3n_{21} - n_{03})(n_{30} + n_{12}) \cdot \left[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2\right] \\
&\quad + (3n_{12} - n_{03})(n_{21} + n_{03}) \cdot \left[3(n_{30} + n_{12})^2 - (n_{21} + n_{30})^2\right]
\end{aligned} \tag{6}$$

3 Watermark Embedding and Extraction

In this section, the watermark embedding and extraction methods are presented, which are applied to the extracted human video objects. Human video object detection is achieved by using a face and body detection module that is based on skin colour distribution, shape information and body location constraints [8]. In the proposed method, a watermark is developed which is invariant to geometric manipulations. As a result, moments continue to exist after geometric distortions. The main idea of the proposed watermarking scheme is to modify video objects in the spatial domain using a predefined function, which determines the moment invariant values in a predetermined space of values.

3.1 Watermark Embedding Method

An overview of the proposed watermark embedding module is depicted in Figure 1. Initially, the human video object *O* is extracted by the human video object detection module. Afterwards, a mapping function *Π* is applied to the extracted human video object *O*, producing a modified human video object $\Pi(\mathrm{O})$, which is used to add noise to the initial video object. In our case study $\Pi(\mathrm{O}) = \log(\mathrm{O})$. Now, let us consider that $\mathrm{M} = \left[\mathrm{m}_1, \ldots, \mathrm{m}_\mathrm{n}\right]^\mathrm{T}$ is the invariant moments vector of the original human video object O. Let also $\mathrm{M}^* = \left[\mathrm{m}_1^*, \ldots, \mathrm{m}_\mathrm{n}^*\right]^\mathrm{T}$ be the invariant moments vector of the watermarked human video object $\tilde{\mathrm{O}}$.

We can choose a function f, which can be any linear or non-linear combination of the invariant moments. In our case study, the function f is expressed as a sum value of the weighted average differences between the moments of the original human video object, M, and the watermarked human video object, M^*:

$$f(M^*, M) = \sum_{i=1}^{n} w_i \left(\frac{m_i^* - m_i}{m_i} \right) \tag{7}$$

These w_i values have been set after several experimental tests for each type of moment. The output of function f is symbolized as factor N (Figures 1 and 2).

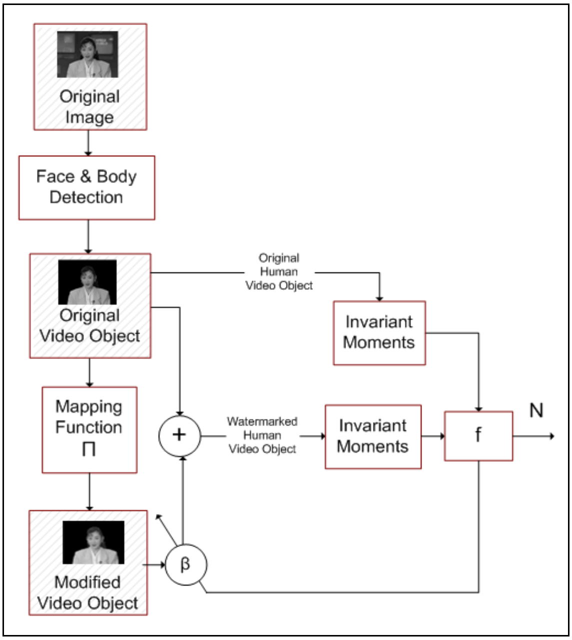


Fig. 1. Block Diagram of the Watermark Embedding Method

The watermarked human video object $\tilde{O}$ can be constructed by adding a perturbation ΔO to the original human video object O: $\tilde{O} = O + \Delta O$. The perturbation ΔO is generated by the following multiplication: $\Delta O = \beta \cdot \Pi(O)$. The weighted factor β is controlled by feedback to ensure $f(M^*, M) \approx 20\%$.

3.2 Watermark Detection Method

In this paper, as the problem of video object watermarking is considered, initially each candidate image (that may contain watermarked video objects) passes through a neural network classifier that detects video objects, which are similar to the watermarked video objects. In this paper, video objects classification is performed using

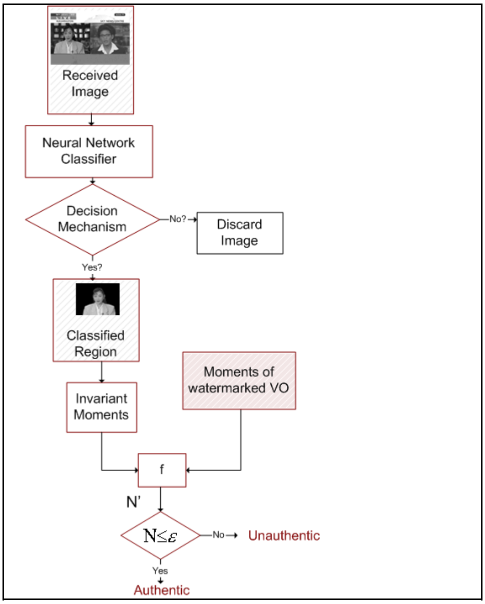


Fig. 2. Block Diagram of the Watermark Detection Method

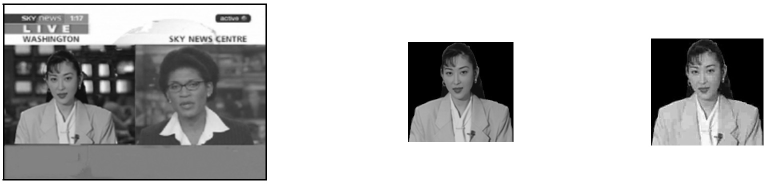
the method described in [9]. In order to efficiently perform the classification task, the neural network is initially trained using blocks of each watermarked video object. The pixels classified to each class constitute the area of the respective video object. However since there is also the case of false detection (detected areas do not really come from any of the respective watermarked video objects), a decision mechanism is incorporated which checks whether the extracted region has size larger than a threshold *T* (procedure continuous) or less than *T* (region is discarded).

Having received a region that the neural network has classified as possible watermarked video object, the detection module estimates the moments of this region. In parallel, the detection procedure receives in its input, the values of the moments of the respective watermarked human video object. Now, the value of factor *N* can be found, counting the mean value of weighted average differences between moments of the watermarked human video object, M^*, and of the candidate region, M'. Authentication of the received human video object can be achieved by checking the validity of the equation $\mathrm{N} \leq \varepsilon$ where ε is the margin of acceptable error between the two video objects. Then, the detection procedure returns either 1, meaning that the candidate video object is the watermarked video object, or 0, meaning that the candidate video object is not the watermarked video object.

4 Experimental Results

In this section, a full experiment is carried out. Let us consider that a frame of the Akiyo sequence is initially analyzed leading to extraction of the 'Akiyo' video object.

This video object is watermarked using the proposed methods and the final content becomes available to the public. Afterwards, a malicious user crops the watermarked human video object (Akiyo) and places it in a different content, as depicted in Figure 3a. Next, the malicious user also performs different types of attacks to the synthetic content producing some final images. Let us now assume that our system receives at its input these final images. At the first step, the neural network classifier extracts a candidate region which is depicted in Figure 3c. In this experiment, threshold *T* has been adjusted to 40%.



a) Watermarked human VO inserted in different background

b) Original human VO

c) Extracted human VO

Fig. 3. Result of the humanVO detection procedure

Next, the watermark detection module is activated and the moments of the candidate region are estimated and they are compared to the moments of the original human video. Several cases are investigated under different types of attacks such as noise addition, compression, filtering, cropping and other image distortions. Results are depicted in Table I where ε has be set equal to 0.01.

It is obvious that the watermarked video object can be detected after most attacks with both categories of moments. As it can be noticed, if the watermarked video object is cropped, performance of the detector deteriorates fast (according to the size of the cropped region). Furthermore, we can see that for a rotation 60° and for jpeg

Table 1. Results of Detecting the Watermarked Akiyo video object After Several Attacks

Attack	$f(M^*,M)=\sum_{i=1}^{N} w_i\left(\frac{M_i^*-M_i}{M_i}\right)$		Watermark Detection	
	HU moments	Zernike moments	HU moments	Zernike moments
Gaussian Filter	0.004023	0.005765	Pass	Pass
Median Filter	0.004023	0.005115	Pass	Pass
JPEG Q=50%	0.001229	0.000547	Pass	Pass
JPEG Q=10%	0.04691	0.000928	Fail	Pass
Rotation -1°	0.000012	0.000004	Pass	Pass
Rotation 5°	0.000456	0.000237	Pass	Pass
Rotation 60°	0.004493	0.000123	Fail	Pass
Scaling 50%	0.002085	0.001034	Pass	Pass
Scaling 110%	0.003790	0.002858	Pass	Pass
Cropping 20%	0.040962	0.050998	Fail	Fail
Flipping	0.002162	0.00124	Pass	Pass

compression with quality 10% only Zernike moments lead to the extraction of the watermark meaning that they can be consider as a more robust watermarking method compared to Hu moments. Cropping 20% the human video object leads to the failure of the detection of the watermarking, even though the neural network classifier found the human video object. It should also be emphasized that the decoder does not need to receive the original video object for detecting the watermark. The decoder only needs to receive the values of the moments of the watermarked video object.

References

1. Kutter, M., Bhattacharjee, S. K., Ebrahimi, T.:Towards second generation watermarking schemes. Proc. IEEE Int. Conf. Image Processing (1999) 320–323.
2. Lin, C. Y. , Wu, M., Bloom, J. A., Cox, I. J., Miller, M. L., Lui, Y. M.: Rotation, scale, and translation resilient watermarking for images. IEEE Trans. Image Processing **10** (May 2001) 767–782.
3. Abu-Mostafa,Y.S., and Psaltis, D.: Image normalization by complex moments. IEEE Trans. Pattern. Anal. Machine Intell. **PAMI-7**, (1985) 46–55.
4. Hu, M. K.: Visual pattern recognition by moment invariants. IEEE Trans. Inform. Theory **8** (1962) 179–187.
5. Zernike, F.: Beugungstheorie des Schneidenverfahrens und seiner verbeserten Form, Phasenkontrastmethode Physica **1** (1934) 689-701.
6. Teague, M.: Image analysis via the general theory of moments,. J. Opt Soc. Am. 70, 8 (1980) 920–930.
7. Alghoniemy, M., and Tewfik, A. H.: Image watermarking by moment invariant. Proc. IEEE Int. Conf. Image Processing. (2000) 73–76.
8. Tsapatsoulis, N., Avrithis, Y., Kollias, S.: Facial Image Indexing in Multimedia Databases. Pattern Analysis & Applications **4** (2001) 93-107.
9. Doulamis, N., Doulamis, A., Ntalianis, K., and Kollias, S.: An Efficient Fully-Unsupervised Video Object Segmentation Scheme Using an Adaptive Neural Network Classifier Architecture. IEEE Transactions on Neural Networks **14** (2003) 616-630.

Cross-Layer Scheduling with Content and Packet Priorities for Optimal Video Streaming over 1xEV-DO*

Tanir Ozcelebi[1], Fabio De Vito[1,3], M. Oğuz Sunay[1], A. Murat Tekalp[1,2],
M. Reha Civanlar[1], and Juan Carlos De Martin[3]

[1] Dep. of Electrical Engineering, Koc University, Sariyer, Istanbul 34450, Turkey
{tozcelebi, fdevito, osunay, mtekalp, rcivanlar}@ku.edu.tr
http://home.ku.edu.tr/~tozcelebi, http://home.ku.edu.tr/~osunay,
http://home.ku.edu.tr/~mtekalp, http://mvgl.ku.edu.tr
[2] Dep. of Electrical and Computer Eng., Univ. of Rochester, Rochester, NY 14627, USA
tekalp@ece.rochester.edu
http://www.ece.rochester.edu/~tekalp/
[3] Dip. Automatica e Informatica, Politecnico di Torino, Torino 10129, Italy
{fabio.devito, juancarlos.demartin}@polito.it
http://media.polito.it/devito,
http://media.polito.it/demartin

Abstract. Maximization of received video quality and application-level service fairness are the two principal objectives of multi-user wireless video streaming. The user and packet scheduling mechanisms employed are the determining factors on the communication system performance and must utilize multiple layers of the OSI protocol stack. The semantic and decodability (concealment related) importance and hence priorities of video packets can be considered at the application layer. In this paper, the use of video content and packet priorities for multi-objective optimized (MOO) scheduling in 1xEV-DO system is introduced. Rate adaptive AVC/H.264 encoding is used for content adaptation and a user with the least buffer fullness, best channel throughput and highest video packet importance is targeted for scheduling. Hence, losses are forced to occur at packets with low semantic/decodability importance. Results show that the proposed framework achieves 1-to-2 dB's better PSNR in high importance temporal regions compared to the state-of-the-art CBR encoding case.

1 Introduction

The increasing bandwidth availability in wireless networks made it possible to distribute multimedia content to mobile users along with classical applications. In this sense, Code Division Multiple Access (CDMA) networks are particularly useful in the case of video transmission, which is quite demanding in terms of network bandwidth. This kind of service in mobile communications requires both high speed and large buffer capacity in handset devices, and the network resource sharing algorithm has to take into account the wide spectrum of receivers logged into the

* This work is supported by 3DTV-NoE funded by the European Commission 6th Framework Information Society Technology Programme and Centro Supercalcolo Piemonte (CSP), Torino, Italy.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 76 – 83, 2006.

© Springer-Verlag Berlin Heidelberg 2006

network, while providing fast access to information content. Quality-of-Service (QoS) is not guaranteed for such applications in most existing systems. Therefore, highly efficient systems that enable high-speed data delivery along with voice support over wireless packet networks are required. Also, there is need for adaptive and efficient system resource allocation methods specific to transmission of such information. Among these methods, opportunistic multiple access schemes [1] in which all system resources are allocated (scheduled) to only one user at a time are known to be optimal due to channel utilization (overall capacity).

In the 1xEV-DO (IS-856) standard [2], opportunistic multiple access is used and all transmission power is assigned to only one user at a time within time slots of length T_s (1.667 ms). The main target is to transmit packet data to multiple users on CDMA/HDR [3] systems at high speed. Adaptive coding and modulation are employed to support various service types (data rates) that can be properly received by a user at all times along the duration of a communication session. It is crucial to choose an appropriate resource (time) scheduling algorithm to achieve the best system performance. Application layer requirements and physical layer limitations need to be well determined, and the scheduler has to be designed accordingly. For example, e-mail and SMS services are tolerant to delay, and intolerant to data loss, while real time streaming applications can tolerate few losses. Hence, cross-layer design is mandatory for video transmission, in order for a scheduling algorithm to be optimal in both physical and application layer aspects.

The state of the art scheduling algorithms for the 1xEV-DO system are maximum C/I (carrier-to-interference ratio) [1] scheduler, also known as the maximum rate scheduler, first in first out (FIFO), proportionally fair (PF) [4] and exponential schedulers [5]. All of these schemes suffer from either application level service fairness or overall channel throughput; hence the average received video quality.

Since our main interest is video streaming rather than a download-and-play solution, video packets that are delivered later than their playout times are discarded at receiver side and are considered lost. Therefore, if the sender detects that the packet will arrive later than its due at the receiver, it can discard (not transmit) the late information at the source side so lowering network congestions.

In video coding, the importance levels of video packets differ from each other due to temporal variance of semantic importance, and also inter and intra frame prediction. The overall user utility can be significantly increased using cross-layer design and appropriate packet priority assignment according to decodability and content (semantic relevance) issues. In this paper, a novel cross-layer multi-objective optimized (MOO) scheduler for video streaming over 1xEV-DO system is presented. The overall channel throughput, individual buffer occupancy levels and contribution of the received network packets in terms of visual quality are simultaneously maximized.

This paper is organized as follows: The scheduling multi-objective optimization (MOO) formulation is outlined in Section 2. The MOO solution methodology employed is explained in Section 3. The experimental results with different settings are given in Section 4, and finally, the conclusions are drawn in Section 5.

2 Problem Formulation

Video contents have not uniform semantic importance within a sequence; since there may be several runs of temporal shots which can be of different interest for different users. It is possible to encode each temporal region at a different bitrate, allowing low-importance frames to be coded at much less bitrate than important ones, thus saving bits for high-importance scenes. In this way, we can obtain better PSNR for semantically important regions. In order to generate this effect, each semantic region has to contain an integer number of GOP's, i.e. Group-of-Pictures. Therefore, GOP size needs to be flexible, and moreover the bitrate control should be able to change its target value for each GOP while encoding the sequence. Packets belonging to the same region have not the same decodability importance; usually, packets coming from I- and P-frames have higher impact on the decoded video if lost, given the possibility of error propagation by means of motion prediction, as described in [6]. The joint importance is computed as the product of the semantic importance level and the quantized decodability importance. Decodability importance is a real number and has to be quantized.

Transmission of video content over low bandwidth channels requires pre-fetching of data stream at the receiver side, so that distortion and pauses caused by buffer underflows or overflows in the duration of video playout can be avoided. This pre-roll (initial buffer) delay can not be excessive for any particular user due to buffer limitations and customer convenience. High visual quality, low pre-roll delay and continuous playout of the content are the most important requirements from a video streaming system, and appropriate scheduling algorithms are desirable.

Both physical layer feedback (C/I ratios) and application layer feedback (decoder buffer level) are used in the proposed framework. In the 1xEV-DO scheme, the back-channel is used to report the current C/I ratio experienced by mobile users, so that the transmitter is aware of the maximum rate that can be achieved for each user within a probability of error range. Channel statistics history is stored and used at the transmitting side for better performance. In our framework, the client buffer occupancy levels are also reported back to the base station.

Assume that there exist K users within the wireless network, demanding videos from the base station with a certain bitrate distribution, $R_V(t)$. Here t $(0 \leq t \leq \infty)$ denotes the discrete time slot index. Our aim is to maximize the overall average channel throughput at each time slot, $R(t)$, while guaranteeing fair and satisfying quality of service for each of these K users. Fairness can be provided by maximizing the buffer levels of individual candidates for scheduling at each time slot. If buffer underflows are inevitable, the video quality can still be protected by careful priority assignment to video packets according to per-packet decodability and semantic importance. In this way, since the video packets with high decodability and semantic importance are transmitted with priority, *packet losses are forced to occur at the less important parts.* Therefore, the group of objective functions to be optimized among users at time t is $\{B_i(t),R_i(t),imp_i(t)\}$, where $B_i(t)$ denotes the buffer fullness level, $R_i(t)$ represents the effective channel throughput, and $imp_i(t)$ is the per-packet importance for user i at time t. The average channel throughput up to time slot t can be calculated as below:

$$\overline{R}(t) = \frac{1}{t} \times \sum_{1 \le i \le k} \sum_{1 \le t' \le t} s_i(t') \cdot R_i(t') \tag{1}$$

where $s_i(t)$ is a binary variable taking the value 1 if user i is scheduled at time slot number t, 0 otherwise. The buffer occupancy level of user i at time t is given by

$$B_i(t) = \max\{B_i(t-1) + T_s \times (s_i(t) \cdot R_i(t) - R_v(t)), 0\} \tag{2}$$

We can also calculate the channel throughput in a recursive manner in terms of previous value as given below:

$$\begin{aligned} \overline{R}(t) &= \frac{1}{t} \times \left((t-1) \times \overline{R}(t-1) + \sum_{1 \le i \le k} s_i(t) \cdot R_i(t) \right) \\ &= \frac{(t-1) \times \overline{R}(t-1)}{t} + \frac{1}{t} \cdot \sum_{1 \le i \le k} s_i(t) \cdot R_i(t) \end{aligned} \tag{3}$$

For large values of t, the first term on the right hand side of the above equation becomes approximately equal to $\overline{R}$(t-1). Then, the throughput enhancement due to scheduling the i^{th} user at time slot t, $\Delta \overline{R}_i$ (t), is calculated as follows:

$$\overline{\Delta R_i}(t) = \overline{R}(t) - \overline{R}(t-1) \cong \frac{1}{t} \cdot R_i(t) \tag{4}$$

Ideally, the server side must schedule the user that experiences the best compromise between the least buffer occupancy level, the best available throughput enhancement and the most important network packet to be delivered. Hence, our optimization formulation for choosing the user to schedule at time slot t is given by

$$\arg\max_i(\overline{\Delta R_i}(t)) = \arg\max_i \left\{ \frac{1}{t} \cdot R_i(t) \right\} \tag{5}$$

$$\arg\min_i(B_i(t)) \tag{6}$$

$$\arg\max_i(imp_i(t)) \tag{7}$$

jointly subject to

$$B_i(t+1) \le BufferSize(i)$$

where *BufferSize(i)* denotes the available decoder buffer size at the i^{th} client. The last constraint is necessary to guarantee that a user whose buffer will overflow after a possible slot assignment is never scheduled. This constraint can indeed cause performance drops in terms of channel capacity especially in the case of maximum rate scheduler, since the user with the highest available rate can not be scheduled for all times.

It is not possible to suggest a direct relationship between the values of instantaneous buffer level and available channel rate for a specific user. In fact, a user's buffer level gives no obvious hint about its current channel condition and vice versa. Therefore, the exhaustive search method for multi-objective optimization given in Section 3 needs to be applied.

3 Multi-objective Optimization (MOO)

For the objective function set $F=\{f_1,f_2,\ldots,f_N\}$, a solution s^* is called globally Pareto-optimal if any one of the objective function values cannot be improved without degrading other objective values. For single objective optimization problems, one can come up with one or more optimal solutions resulting in a *unique* optimal function value, and a Pareto-optimal solution is also a globally optimal solution. In contrast, this uniqueness of the optimal function value is not valid for multi-objective optimization (MOO) problems since two or more of the objective functions may be either conflicting or uncorrelated. Hence, it can be the case that there exist many Pareto-optimal solutions and one has to discriminate between these solutions to determine which one is better and to come up with a best compromise solution. For this, one needs to determine the relative importance of objective functions. In case of equally important objectives, the individual objective values need to be rescaled to an appropriate range in order to compensate for range differences as follows:

$$f_{i,scaled} = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}} \tag{8}$$

In our method, the throughput enhancement, the decoder buffer occupancy level and the per-packet importance are normalized to take real values between 0 and 1 as shown in Fig. 1 for the two dimensional case, which is also described in [7].

In solution of such problems, an *infeasible* point that optimizes all of the objective functions *individually* is called the utopia point. The utopia point, *U(t)*, on the throughput-buffer-importance space is set as follows:

$$U(t) = \left(\overline{\Delta R}(t)_{\max}, B(t)_{\min}, imp_i(t)\right) \tag{9}$$

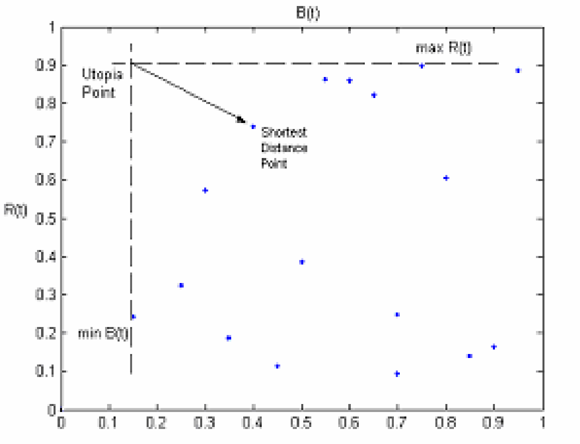


Fig. 1. The proposed algorithm schedules the user whose corresponding point is closest to the utopia point

The best compromise optimal solution is found as the *feasible* point that is closest to the utopia point in the Euclidian-distance sense. A more detailed explanation of multiple-objective optimization (MOO) techniques used in the literature can be found in [8]-[9].

4 Experimental Results

We encoded a 2250-frame test sequence (part of a soccer match), whose time duration is 90 seconds. Semantically more important regions are coded at a higher rate than low-importance GOP's, at bitrate ratios 1-to-2 and 1-to-3 as shown in Table 1 and Table 2, resulting in an average bitrate of 100 kbps in each case. The decodability importance has been quantized using two levels.

The resulting bitstreams are fed into the scheduler using the product of semantic and decodability importance indicators. Losses are introduced only by late packet delivery. Twelve users require the same video at random times within a time period of 5 seconds. Packets are ordered on a GOP-basis at the source side, according to their importance. In this way, we transmit important packets of each GOP first, and packets discarded due to late delivery will be concentrated in easily-concealed and low-importance regions. The maximum allowed initial buffering time is set to 5 seconds. Alternatively, a user stops pre-fetching after half of the decoder buffer is already full, where the decoder buffer size is set to 1 Mbits.

Results for overall PSNR obtained are shown in Table 1. In the 1-to-1 rate allocation case, the whole video is encoded and played at constant bitrate avoiding peaks in the rate distribution. Here, the network is highly loaded with 1200 kbps (12×100 kbps) peak rate almost all the time during transmission, causing excessive

Table 1. Packet loss rates and overall PSNR for the test sequence, using 2 levels of semantic importance and 2 levels of decodability importance

	1 to 1 rate allocation		1 to 2 rate allocation		1 to 3 rate allocation	
User	PLR (%)	PSNR (dB)	PLR (%)	PSNR (dB)	PLR (%)	PSNR (dB)
1	0.89	34.24	0	34.16	1.49	33.54
2	0.14	34.45	0	34.16	1.97	33.32
3	0.90	34.24	1.18	33.85	2.01	33.45
4	0	34.504	0.96	33.90	0.95	33.58
5	1.85	34.24	1.17	33.87	3.53	33.24
6	1.70	34.09	0.17	34.13	0.75	33.65
7	0.11	34.47	0.17	34.13	0.81	33.59
8	0.20	34.40	0.61	34.03	0.75	33.71
9	0.35	34.39	0	34.16	0.34	33.77
10	0.87	34.34	0	34.16	1.21	33.55
11	0.86	34.27	1.18	33.86	1.14	33.60
12	0	34.504	0.96	33.91	0.47	33.75

Table 2. Packet loss rates and PSNR for the high importance region of the test sequence, using 2 levels of semantic importance and 2 levels of decodability importance

	1 to 2 rate allocation		1 to 3 rate allocation	
User	PLR (%)	PSNR (dB)	PLR (%)	PSNR (dB)
1	0	36.03	1.84	35.99
2	0	36.03	1.74	36.00
3	0.23	35.97	2.21	35.96
4	0	35.97	0.19	36.45
5	0.77	35.77	4.35	35.52
6	0.22	35.92	0.91	36.19
7	0.12	35.97	0.14	36.41
8	0.09	35.96	0.93	36.29
9	0	36.03	0.41	36.35
10	0	36.03	0.50	36.38
11	0.40	35.94	1.41	36.12
12	0	35.97	0.58	36.32

packet loss rate. On the other hand, in the 1-to-2 ratio case, the rate distribution is not uniform. Considering that the users are accessing the network at random times, some of the users will be draining data at 122 kbps, while others are streaming at 61 kbps, smoothing out the peak required transmission rate values. Hence, this unequal rate distribution is actually useful for reducing packet losses. In the 1-to-3 rate ratio case, the bitrates of the semantically important segments are themselves too high, resulting in packet losses higher than in the case of 1-to-2 ratio case.

The overall PSNR of the sequences decreases as the gap of semantic importance increases along segments. This is natural due to the effect of non-linear mapping between bitrate and PSNR. Here, the low importance parts have much lower PSNR than the 1-to-1 case, while high importance levels gain 1 or 2 dB's with respect to the same level. If we focus our attention only on the high importance parts, the PSNR increases with wider gap in the rate allocation of different regions.

5 Conclusions

In this paper, we proposed a novel cross-layer optimization technique for determining the best allocation of channel resources (time slots) across users over 1xEV-DO wireless channels. The novelty of this framework comes from the usage of decodability and semantic importance feedback from the application layer to the scheduler. The modifications to the H.264 codec have been described as well as the optimized scheduling algorithm. Network simulations show that noticeable improvements can be obtained with respect to the scheduler which does not consider packet importance, especially under strict requirements such as very short pre-roll delays. Experimental results show that, this approach ensures higher video PSNR with respect to constant bitrate coding. Furthermore, to better simulate the actual user behavior, we introduced random initial access times for users. As a result, received video PSNR was further improved.

References

1. Knopp, R., Humblet, P. A.: Multiple accessing over frequency selective fading channels: Proceedings of IEEE PIMRC Toronto, Canada (1995)
2. TIA/EIA IS-856-2: Cdma2000 High rate packet data air interface specification (2002)
3. Bender, P., Black, P., Grob, M., Padovani, R., Sindhushayana, N., Viterbi, A.: Cdma/hdr: A bandwidth efficient high-speed wireless data service for nomadic users: IEEE Communications Magazine 38 (7) (2000) 70-77
4. Jalali, A., Padovani, R., Pankaj, R.: Data throughput of cdma-hdr: A high efficiency high data rate personal communications system: IEEE 51st Vehicular Technology Conference, Tokyo, Japan (2000)
5. Shakkottai, S., Stolyar, A.: Scheduling algorithms for a mixture of real-time and non-real-time data in HDR: Proc. 17th International Teletraffic Congress (ITC-17), Salvador de Bahia, Brazil (2001)
6. De Vito, F., Quaglia, D., De Martin, J. C.: Model-based distortion estimation for perceptual classification of video packets: Proceedings of IEEE Int. Workshop on Multimedia Signal Processing (MMSP) 1, Siena, Italy (2004) 79-82
7. Ozcelebi, T., Sunay, O., Tekalp, A. M., Civanlar, M. R.: Cross-layer design for real time video streaming over 1xEV-DO using multiple objective optimization: to appear in IEEE GlobeCom (2005)
8. Papadimitriou, H., Yannakakis, M.: Multi-objective query optimization: Proceedings of Symposium on Principles of Database Systems (PODS), California (2001) 52-59
9. Lim, Y.-il, Floquet, P., Joulia, X.: Multiobjective optimization considering economics and environmental impact: ECCE2, Montpellier (1999)

Video Rendering: Zooming Video Using Fractals

Maurizio Murroni and Giulio Soro

Department of Electrical and Electronic Engineering, University of Cagliari,
P.zza d' Armi, Cagliari 09123, Italy
{murroni, giulio.soro}@diee.unica.it
http://mclab.diee.unica.it

Abstract. Slow motion replay and spatial zooming are special effects used in video rendering. Already consolidated as commercial features of analog video players, today both these effects are likely to be extended to the digital environment. Purpose of this paper is to present a technique combining fractals (IFS) and wavelets to obtain a subjectively pleasant zoom and slow motion of digital video sequences. Active scene detection and post processing techniques are used to reduce computational cost and improve visual quality respectively. This study shows that the proposed technique produces better results than the state of the art techniques based either on data replication or classical interpolation.

1 Introduction

Among all possible interactive applications, widely used in classic analog video reproduction, slow motion replay is one of the most expected to be extended to the digital formats. Slow motion is a commercial feature of home video players, but also, a special effect used in the video production field. In a analog framework, given a video sequence with a fixed frame rate f, classical slow motion effect is obtained reducing the frame rate to $f^{'} < f$, so that a frame remains visible for a time proportional to slow motion factor. At present, commercial digital video players allow users to browse a video sequence frame by frame, or by chapter selection with prefixed indexes. Slow motion replay is achieved by reducing the frame rate display or keeping it constant and inserting within the sequence additional intermediate frames generated by linear or cubic interpolation. A major drawbacks of these approaches are that interpolation for slow motion replay yields to a "fading" effect between frames, whereas frame replication creates a "jerky" distortion, both resulting in low motion quality for the human visual system. Similar issues arise in image plane if pixel replication or interpolation is used to perform spatial zoom. Several works were published in the past on this topic. One of most interesting algorithm, based on motion estimation techniques, was developed at the B.B.C. labs by G.A. Thomas and H.Y.K. Lau [1], [2]: the frame was divided into partially overlapped blocks. By means of Fourier analysis, a phase correlation was performed between corresponding blocks belonging to adjacent frames. Moving vectors were identified and interpolated to generate missing frames. The main weakness of this technique was the inability to deal with the case of motion detection failure. This could occur due to the presence of high speed

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 84 – 91, 2006.
© Springer-Verlag Berlin Heidelberg 2006

movement in the scene, so that the motion estimation algorithm was unable to find a good approximation of the movement for each block of the scene. Therefore, in presence of high speed movement in the sequence, the effectiveness of the latter method decreased.

In this work, we propose an alternative post processing scheme that combines the properties of fractal representation of a video sequence with motion detection techniques and wavelet subband analysis. The aim is to overcome the above mentioned annoying effects and obtain a pleasant zoom and slow motion.

In literature fractals on image applications were proposed to achieve data compression exploiting pseudo-self-similarity inside natural images [3], [4]. But the potentiality of fractals is not limited to compression. The properties of fractal coding allow expanding a multidimensional signal (e. g. image and video sequences) along its dimensions.

One of the major weaknesses of the fractal representation of a signal is the high computational complexity of the encoding process. The computational load and the processing time increases for signals of higher dimension (1D, 2D, 3D…). This is due to the fact that there are more data to be processed increasing the dimension of a signal. Furthermore, as it will be explained in detail later through the manuscript, the main idea of fractal encoding is to look for similarities of blocks using affine transforms that are much more for a multidimensional signal respect to the monodimensional case, so that a *best match algorithm* leads to a great time consuming process for multidimensional data sets. Several methods have been proposed in literature to speed up the fractal coding process [5]: a class of proposed solutions is based on wavelet subband analysis. Due to their orthogonal and localization properties, wavelets are well suited (and extensively adopted) for subband data analysis and processing. The proposed algorithm exploits these features performing the fractal coding of the subbands extracted by means of the wavelet analysis, with particular attention to the frequency distribution of the coefficients, in order to perform an efficient coding process. Moreover, to further reduce the high computational cost of fractal encoding an active scene detection is used so as to perform three-dimensional fractal coding only in high information areas (moving areas), whereas static zones are coded using two dimensional coder. As suggested in [6], for two dimensional fractal coding, some post-processing techniques, extended to the three dimensional case, are used to improve overall visual quality. Results show that the quality achieved is higher if compared to the state of the art techniques, as well as time coding reduction makes the method suitable for interactive multimedia applications.

The rest of the paper is organized as follows: in section II the proposed method is presented; experimental results are provided in section III and conclusion are finally given in section IV.

2 Proposed Method

Within a framework of interactive applications, the user should select a precise scene of interest (i.e., a sub-sequence corresponding to the desired time interval) to be spatially zoomed and replayed in slow motion. Let the selected sub-sequence be composed by M frames. At first, being the computational complexity of the fractal encoder strictly proportional to the amount of data to be processed, frames are grouped into packets (GOP) with length *N*. *N* is chosen according to the temporal activity of the sequence, so that higher values can be selected for slowly changing scenes without a significant time processing increase. Each GOP is treated as a single unit to be coded. The drawback of this packetization process is that it introduces a discontinuity along the temporal axis. To limit this effect, time overlapping is used: each GOP is coded having as a boundary condition the motion information of the previous one. Within each GOP an active scene detector is used to find the "active object" so that a group of three dimensional blocks is extracted. Each frame is divided into tiles of $M \times M$ size. The *Minimum Square Error* (MSE) among corresponding tiles belonging to different frames is computed. If the MSE is higher than a prefixed threshold, tiles are grouped to form a three dimensional block. The threshold is adaptively evaluated by averaging the MSE over all tiles composing the GOP. The set of the so extracted blocks defines the active object, the remaining blocks constituting the "background". The active object is suited to be coded with a full threedimensional fractal coder whereas the static background is processed with a two-dimensional one. Fractal coding is performed according to the IFS theory [3]: at first, data are partitioned into *range* and *domain blocks*; then, a *domain pool* is created by means of domain blocks and their contractive affine transforms. Each range block is then compared to the elements composing the domain pool by means of the MSE and a set of correspondences among range blocks, domain blocks and affine transforms (i.e., the fractal code) is created.

To enhance fractal coding, Overlapped Range Blocks (ORB) technique [6] is used (see fig. 1). ORB coding is extended to the three dimensional case for the active object coding. Background is encoded with a two-dimensional fractal code, since data does not change, on the temporal axis, for background blocks.

To speed up the fractal coding process a wavelet subband analysis and coefficient classification [9] is executed both for background and active object. For the active object a three dimensional wavelet subband analysis is computed. For the entire low pass component a fractal code is then extracted using ORB partitioning. For the high-pass components, the following coefficients classification procedure is performed: let S_m be the m-th subband; we denote by $\left\{x_i^m\right\}$ the wavelet coefficients of S_m and by $p^m\left(x\right)$ the histogram of $\left\{x_i^m\right\}$. In $p^m\left(x\right)$, starting from the maximum $x_{\max}$ and moving to the tails of the distribution (see Fig. 4), two thresholds are identified, that is $t_1^m, t_2^m : \int_{t_1}^{t_2} p^m\left(x\right)dx = K, \quad K \in \left(0,1\right]$. These thresholds identifies the wavelet coefficients constituting the active zone for S_m, that is $S_m^{az} = \left\{\forall\, x \in \left\{x_i^m\right\}, x \notin \left[t_1^m, t_2^m\right]\right\}$ In other words, an active zone is composed by those coefficients located on the distribution's tails identified by the above thresholds.

After the classification process, a binary-value mask, indicating the position of active zone coefficients within the subband, is extracted. Those coefficients that do not belong to an active zone are discarded, while the S_m^{az} coefficients are ORB partitioned and then fractal encoded.

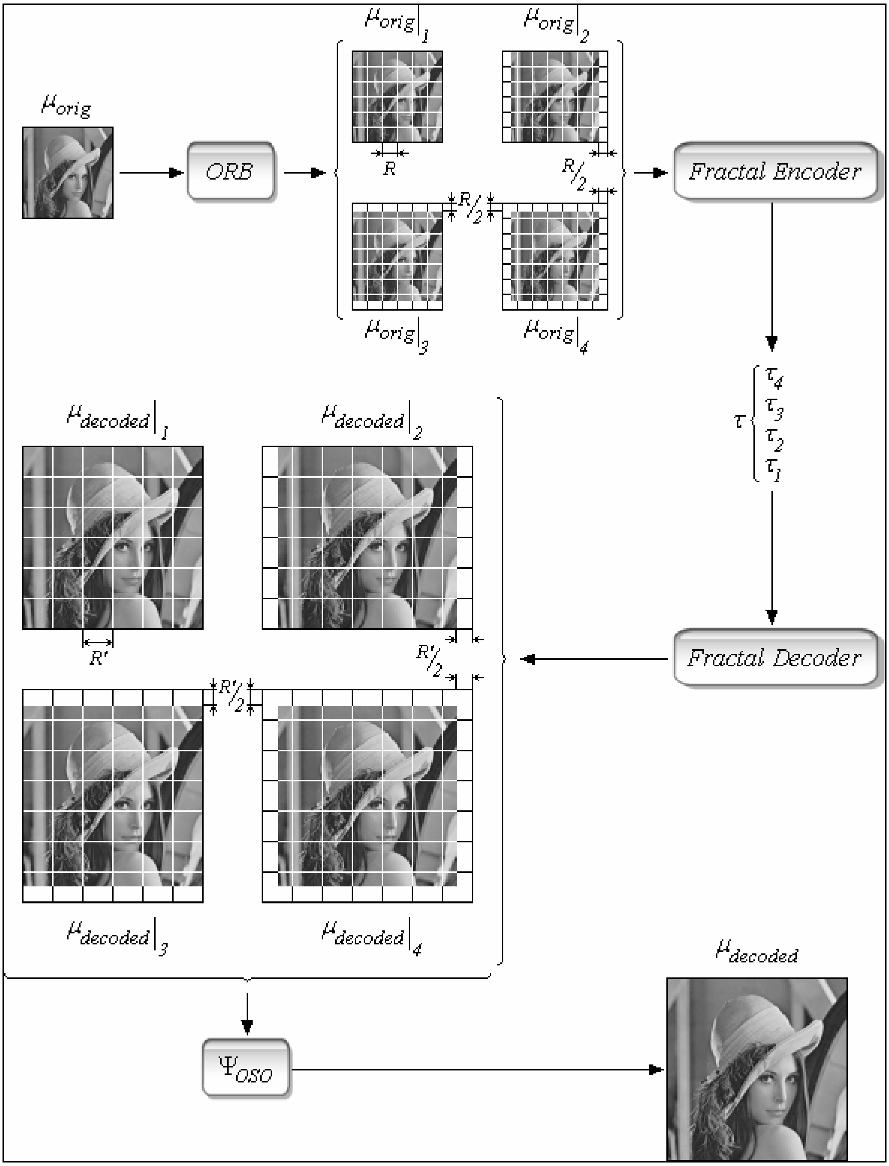


Fig. 1. ORB partition and OSO filtering processes

The K parameter is unique for all the subbands and controls the speed up, and on the other hand, the accuracy of the fractal coding process; higher values of K correspond to higher speed up factors, but also turn out in lower final visual quality achieved.

An additional advantage in terms of time saving of wavelet analysis is the "parallelization" of the entire process that increases the speed in a multithreaded environment.

At decoding time, the inverse process is applied, and the fractal zoom is performed. Since the extracted fractal code is resolution independent, during the decoding process an expansion can be performed independently along each dimension [10]. A three-dimensional (i.e., spatial and temporal) expansion of the active object and two-dimensional spatial zoom (i.e., frames of bigger size) of the background are performed.

After the inverse wavelet transformation, an Ordered Square Overlapping (OSO) [6] post-process, and its direct extension for the three dimensional case called Ordered Cube Overlapping (OCO), are applied to the partition created by the ORB code. Combined ORB code and OSO/OCO filtering enhance visual quality performance of fractal, by coding reducing blocking artifacts generated by the block based nature of the IFS approach.

Finally, an active scene merging and a packet merging processes are applied to release the desired output video sequence. The architecture of the whole process is shown in fig. 2.

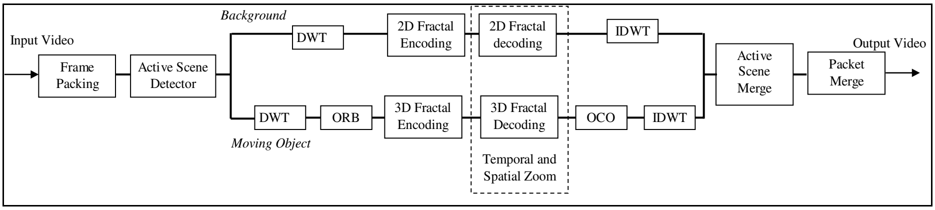


Fig. 2. Sketch of the proposed algorithm

3 Results

We compared the proposed technique with classical frame replication and interpolation considered as benchmark techniques. Several tests were carried out on video sequences with different temporal activity. Test sequences were *Silent*, *Miss America, Stefan, Carphone, Coastguard* and *Mobile* in CIF format. To obtain numerical results, the original sequences were downsampled both in time and image plane. This downsampled version of the original sequence was the input video for the entire process, the aim of which was to obtain again a sequence of the same size of the original. A measure of the overall visual quality achieved was obtained by comparing the expanded sequence to the original video. During the test procedure zoom in both spatial and temporal dimension was performed. Spatial expansion of the sequences was limited to 2x factor corresponding to a double size of the pictures, whereas high slow motion was performed extending the temporal expansion up to 8x factor. This choice was driven by the fact that, especially in the movie entertainment field, the importance of obtaining elevated slow motion factors is prominent. Basically, this is because spatial zooming is limited by the display size used for reproduction. Therefore, much more attention was dedicated during the experiment at the performance of the method in terms of pleasant and "fluid" slow motion reproduction. Moreover, two dimensional fractal zooming features and performance have been extensively investigated and compared to classical interpolators by means of both subjective and objective quality measurements [10].

To evaluate the visual quality achieved, we refer to measures for objective performance assessment defined by [11] and successively included in ITU Recommendations [12, 13]. To extract the performance metrics we deployed the Video Quality Metric (VQM) software developed by the ITS-Video Quality Research project [14]. All the tests performed on the different test sequences produced similar outcomes. Therefore, for the shake of concision, we reported here only a subset of results that were relevant to the *Silent* and *Mobile* sequences and that we reported in table I and II. We considered for the expansion, 64 frames for the sequence *Silent* at 15 frame/s rate and 128 frames for the sequence *Mobile* at 30 frame/s rate, both corresponding to approximately 4 seconds of the video scene.

Table 1. Evaluation of visual video quality for the Mobile sequence (Spatial Zoom: 2x; Slow Motion Factor: 8x)

Quality Metric (normalized)	Fractal Exp.	Frame Replica	Frame Interp.
Jerkiness	38%	44%	45%
Blurring	82%	82%	83%
Block Distortion	19%	22%	19%
Error Blocks	0%	3%	0%

Table 2. Evaluation of visual video quality for the Silent sequence (Spatial Zoom: 2x; Slow Motion Factor: 8x)

Quality Metric (normalized)	Fractal Exp.	Frame Replica	Frame Interp.
Jerkiness	28%	34%	40%
Blurring	75%	77%	79%
Block Distortion	14%	20%	17%
Error Blocks	0%	2%	0%

Table 3. Evaluation of visual video quality for the Silent sequence (Slow Motion Factor: 2x)

Quality Metric (normalized)	Fractal Exp.	Frame Replica	Frame Interp.
Jerkiness	11%	21%	17%
Blurring	8%	10%	7%
Block Distortion	42%	59%	45%
Error Blocks	0%	4%	3%

Table 4. Evaluation of visual video quality for the Silent sequence (Slow Motion Factor: 4x)

Quality Metric (normalized)	Fractal Exp.	Frame Replica	Frame Interp.
Jerkiness	12%	33%	21%
Blurring	13%	19%	16%
Block Distortion	49%	79%	51%
Error Blocks	3%	18%	13%

Table 5. Evaluation of visual video quality for the Mobile sequence (Slow Motion Factor: 2x)

Quality Metric (normalized)	Fractal Exp.	Frame Replica	Frame Interp.
Jerkiness	18%	22%	19%
Blurring	7%	10%	8%
Block Distortion	55%	57%	45%
Error Blocks	7%	13%	10%

Table 6. Evaluation of visual video quality for the Mobile sequence (Slow Motion Factor: 4x)

Quality Metric (normalized)	Fractal Exp.	Frame Replica	Frame Interp.
Jerkiness	21%	39%	22%
Blurring	21%	24%	23%
Block Distortion	59%	79%	61%
Error Blocks	11%	18%	12%

4 Conclusions

We presented here a technique combining fractals and wavelets to obtain a subjectively pleasant zoom and slow motion of digital video sequences. The proposed method achieves high slow-motion ratios with a pleasant visual quality in both time and image plane reducing space and temporal distortions introduced by classical techniques. Computational cost, the main problem of fractal coding, has been reduced using a joint motion detection and wavelet subband analysis approach.

References

1. Thomas, G. A.: Distorting the time axis: motion compensated image processing in the studio. IBC’88 IEE Conference Publication no 293 (1988) 256-25.
2. Thomas, G. A., and Lau, H. Y. K.: Generation of high quality slow motion replay using motion compensation. Conference Proceeding, IBC (1990) 121-125.

3. Barnsley, S., and Demko, M. F.: Iterated function systems and the global construction of fractal. Proc. Royal Soc. London, vc A399 (1985) 243-275.
4. Jaquin, A. E.: Image coding based on a fractal theory of iterated contractive image trasformation. IEEE Trans. on Image Processing **1** no. 1 (1992) 18-30.
5. Polvere, M., and Nappi, M.: Speed-Up in Fractal Image Coding: Comparison of Methods. IEEE Trans. on Image Processing **9** no. 6 (2000) 1002-1009.
6. Reusens, E.: Overlapped Adaptive Partitioning for Image Coding Based on Theory of Iterated Function Systems. Proc. IEEE ICASSP Adelaide, Australia **5** (1994) V/569-V/572.
7. Barthel, K. U., and Voye, T.: Three-Dimensional Fractal Video Coding. Proc. IEEE ICIP-95 Washington, D.C, (1995) III 260-263.
8. Barthel, K. U., Voye, T., and Ruhl, G.: Combining Wavelet and Fractal Coding for 3-D Video Coding. Proc. IEEE ICIP-96 Lausanne **1** (1996) 181-185.
9. Ancis, M., and Giusto, D. D.: Image data compression by adaptive vector quantization of classified wavelet coefficients. Proc. IEEE PACRIM Conference Victoria, Canada, (1997) 330-333,
10. Polidori, E., Dugelay, J-L.: Zooming using Iterated Function Systems. NATO ASI Conference on Fractal Image Encoding and Analysis, Trondheim (1995).
11. ANSI T1.801.03 – 1996: American National Standard for Telecommunications – Digital Transport of One – Way Video Signals – Parameters for Objective Performance Assessment. Alliance for Telecommunications Industry Solutions, 1200 G Street, N. W., Suite 500, Washington DC.
12. ITU-T Recommendation J.144R: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.
13. ITU-R Recommendation BT.1683: Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference.
14. Video Quality Research project: http://www.its.bldrdoc.gov/n3/video/Default.htm

Progressive Mesh-Based Motion Estimation Using Partial Refinement

Heechan Park, Andy C. Yu, and Graham R. Martin

Department of Computer Science,
University of Warwick,
Coventry, United Kingdom
heechan@dcs.warwick.ac.uk

Abstract. A technique for performing progressive mesh-based motion estimation in a layered fashion is presented. Motion compensation based on image warping provides a block prediction free of block artifacts. The smooth prediction can be used to identify motion-active regions by comparing with the reference frame and generate a partial denser mesh, thus forming layers of mesh. This approach provides a hierarchical partial refinement according to motion activity without additional cost. Experimental results indicate that the technique shows improvement over a single-layered uniform mesh and advantages over block-based techniques, particularly in scalable and very low bitrate video coding.

1 Introduction

Block matching motion estimation forms an essential component of inter-frame coding in many video coding standards. The block matching algorithm adopts a translational motion model, but this is intrinsically limited when representing real world motion. In order to cope with complex motion such as rotation and zooming, deformable mesh-based algorithms have been proposed. In general, a mesh consists of polygonal patches, and a spatial transformation function is applied to map the image into a new coordinate system.

Mesh-based motion estimation can be divided into two categories, defined by whether motion is estimated in the forward or backward directions. In backward motion estimation, a mesh is applied to the current frame and deformations are estimated from the current to the reference frame. Forward methods operate in the opposite manner. Backward motion estimation is widely used because of its relative simplicity and the lower computational requirement of the mapping process. Forward methods can provide adaptive deformation of the patch structure to track feature changes in the image, but at the expense of complexity. Mesh-based techniques can be further categorized depending on whether a regular or irregular mesh is employed. A regular mesh consists of uniform patches. An irregular mesh is generated according to the image content using Delaunay or Quadtree methods, and where patch size varies with intensity gradient or motion activity. Generally, a regular mesh is associated with backward estimation. The irregular mesh is coupled with forward estimation to avoid transmitting a large

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 92–98, 2006.
© Springer-Verlag Berlin Heidelberg 2006

overhead for the patch structure. The latter provides better performance than a regular mesh. However it is not popular in real applications due to the high complexity of the forward method or associated overhead for transmitting node positions. In this paper, we present a regular mesh technique with backward estimation that features the advantages of an irregular mesh.

2 ME / MC with Spatial Transform

Motion estimation (ME) and motion compensation (MC) based on a triangular mesh, partitions the image into a number of triangular patches where the vertices are denoted as grid points. Using mesh refinement (Sec. 2.2), the displacement of each grid point is estimated and represented by a motion vector(MV). The displaced grid points define a deformed mesh that describes the underlying motion. The deformed mesh of the reference frame is obtained from estimating the displaced position of the mesh vertices in the current frame. Motion compensation proceeds by retrieving six affine parameters from the displacements of the three vertices of each triangular patch, and synthesizing patch content using a warping operation defined by the six parameters.

2.1 Affine Transform

An affine transform models translation, rotation, and scaling of a patch in the current frame to the corresponding distorted patch in the reference frame. This transformation is represented by six parameters. An intensity value of pixel (x, y) in the i th synthesized patch $\hat{P}$ in the predicted frame K is given by

$$\hat{P}_i^k(x, y) = P_i^{k-1}(f_i(x, y)) \tag{1}$$

where the affine transform $f(\cdot)$ of the patch is given by

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

where (x', y') and (x, y) denote positions in the reference frame and the current frame respectively. There is a one-to-one correspondence between vertices in the current frame and the reference frame, and therefore the six parameters, $a_1, a_2, a_3, b_1, b_2, b_3$, are obtained by solving equations provided by the motion vectors at the vertices and pixels within corresponding patches can be interpolated accordingly.

2.2 Mesh Refinement

Mesh refinement refers to modification of the grid point locations so that the image intensity distribution within any two corresponding patches in the current and reference frame match under an affine transformation. Finding the optimum combination of each grid point that minimizes the difference between the current

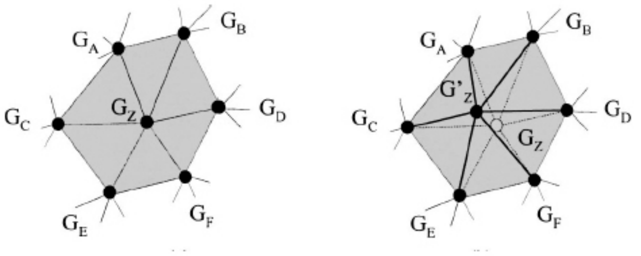


Fig. 1. Hexagon-based mesh refinement: (left) before, (right) after refinement

and reference frame for all possible combinations of grid point locations is not feasible in practice. Instead, finding a sub-optimum combination by successively visiting each grid point of the mesh and moving the grid point to a new position within range, thus preserving mesh connectivity and minimizing matching error locally has been developed by Nakaya et al. [1].The ME of the grid points proceeds with iterative local minimization of the prediction error to refine the MV as below.

While keeping the location of the six surrounding grid points, $G_A \sim G_F$ fixed (Fig.1), G_Z is moved to an adjacent position G'_Z. This is repeated, and for each move, the six surrounding patches inside the hexagon are warped and compared with the current patch by the mean absolute difference. The optimum position of G'_Z is registered as the new position of G_Z. This procedure is repeated for each grid point until all the grid points converge to either local or global minima.

3 Progressive Motion Estimation

A block-based model leads to severe block distortions while the mesh-based method may cause warping artifacts. In terms of prediction accuracy, the mesh-based model can give a more visually acceptable prediction, particularly in the presence of non-translational motion. The complex motion modeling and block artifact-free characteristic of warping enables identification of the approximate difference region between successive frames using the motion vectors alone. This does not appear to have been explored, and is the motivation behind our proposed algorithm.

As mentioned, regular / backward mesh ME is a popular choice due to its relative simplicity and lower computational cost. However, a uniform patch structure cannot accommodate non-stationary image characteristics nor cope with exact object boundary representation. This is addressed by employing a hierarchical structure using a Quadtree[2]. A mesh of a different density is applied according to motion activity, and this allows more accurate motion modeling where needed. However, the technique requires additional bits to indicate the motion active region and has constraints on the flexible movement of the grid points, which add complexity to the refinement process. We propose a progressive mesh refinement method that is adaptable to motion activity in a layered fashion, which overcomes the limitation above.

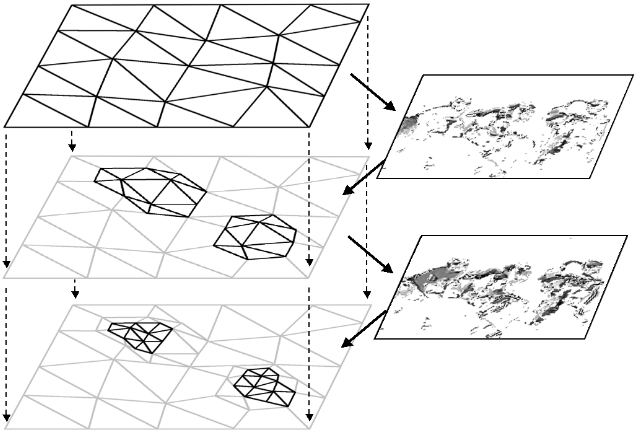


Fig. 2. Progressive mesh refinement layers and approximated frame difference maps

Fig.2 illustrates progressive motion estimation in terms of layer and an approximated frame difference between layers to identify motion-active regions and generate the partial mesh of the next layer. A global motion estimation is performed on the top layer while the next layers concentrate on the finer motion activity. A similar approach was suggested in [3]. While our technique provides partial refinement by employing a layered mesh topology without overhead, Toklu et. al. focus on the hierarchical mesh structure to the entire image in one layer, resulting in higher bit rates. Fig.3 shows an example of an approximated difference map. The actual frame difference between the current and reference frames,(a), is quite well approximated in (c). The approximate difference, (c), is obtained from subtracting the (warped) predicted reference frame (b) from the current frame.

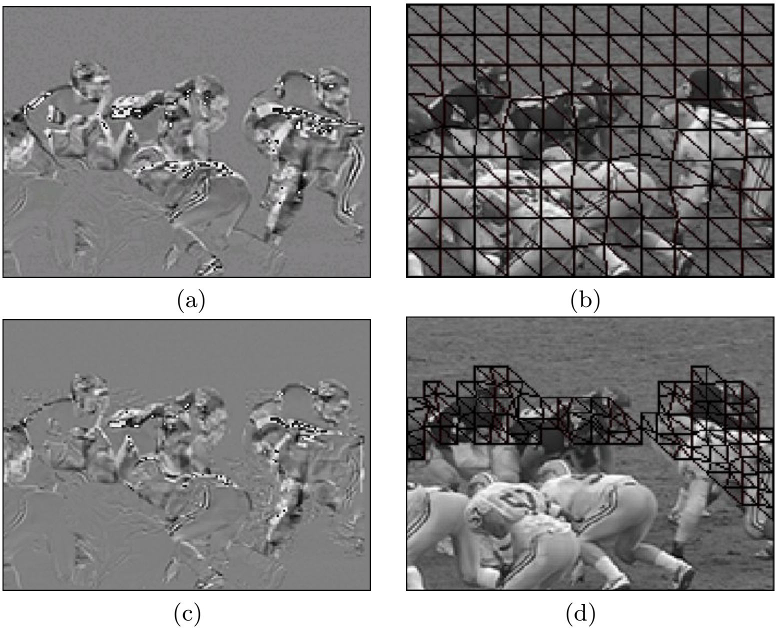
The technique proceeds as follows. Firstly we apply mesh refinement for a coarse prediction with a uniform mesh and synthesize an image from the resulting MVs as in the standard mesh-based algorithm. We then identify regions where motion discontinuities exist from the difference map, that is the variance of the patch difference,v_P is greater than the variance of the frame difference, v_F. The variance of the frame difference is given by

$$v_F = \frac{1}{M \cdot N} \sum_{j=1}^{M} \sum_{i=1}^{N} (f(i,j) - \bar{f})^2 \tag{3}$$

where M and N are the frame dimensions and $\bar{f}$ denotes the mean frame difference. The variance of the patch difference is given by

$$v_P = \frac{1}{K} \sum_{i=1}^{K} (f_P(i) - \bar{f_P})^2 \tag{4}$$

where K refers to the number of pixels in the patch.



(a)

(b)

(c)

(d)

Fig. 3. Frame difference and partial refinement: (a) difference between current and reference images, (b) deformed regular mesh, (c) difference between frame synthesized by (b) and current frame, (d) partial refinement

In the next step, a finer regular mesh is applied to the regions containing motion discontinuities, as depicted in Fig.3 (d). Consequently we have layers of mesh, a coarse mesh that covers the entire image and a denser partial mesh applied to the moving regions only. This allows a hierarchical refinement without the explicit overhead and constraints on movement of the grid points.

4 Experimental Results

The algorithms were evaluated using the QCIF resolution test sequences “Crew”, “Football”, “Ice”, “Suzie” and “Stefan”. According to our experiments, using two layers of mesh with patch sizes of 16×16 and 8×8 show the best performance in QCIF resolution in terms of rate-distortion. The hexagonal matching algorithm[1] is applied with a search range of ± 7 pixels for the first layer and ± 3 pixels for the second layer which preserves mesh connectivity. More grid points in the intial layer do not necessarily lead to either better motion active region identification or a better reconstruction quality when bitrate is considered. This is due to the grid point connectivity constraint that prevents effective estimation when an over-dense mesh covers occluded / discovered areas, and of course the increased number of motion vectors. In this sense, the content-based mesh provides an advantage(see Sec. 5). The motion field is initialized with zero-motion and iteration starts with the grid point closest to the image centre. A

Table 1. Experimental Result: bitrate for MV and PSNR (0.2 bpp)

	One Layer(HMA)		Two Layer	
Sequence	MVs	PSNR	MVs	PSNR
Crew	532	30.93	1190	30.95
Football	645	22.21	1149	22.32
Ice	415	27.11	913	27.68
Susie	349	37.93	719	38.15

hexagon that contains at least one patch overlapping with the identified regions is included in the partial refinement process. Note there is no need to transmit the region information as the region can be identified using the MVs transmitted in each layer. Motion vectors are differentially encoded using Exp-Golomb. Wavelet coding is applied to the displaced frame residue. In Table.1, the left column shows the performance of the single-layered mesh refinement and the right column represents the performance with an additional layer. The overall performance is improved in all test sequences at a fixed bitrate (0.2 bpp). The poor improvement in the Crew sequence can be accounted for by frames containing a flashing light which is more efficiently compressed with residual coding.

5 Future Work

Scalable video coding utilizing the wavelet transform applied to both the spatial and temporal domains (3D-DWT) is of current interest in video coding[4]. The mesh-based ME/MC scheme exhibits several merits when deployed in the wavelet coder[5]. The mesh-based estimation can provide scalable coding naturally by controlling the number of grid points or controlling the number of layers in our algorithm. Also, the unique trajectory of each pixel in a mesh-based motion model overcomes the appearance of so-called “multiple/unconnected” pixels occuring in areas not conforming to the rigid translational model. S. Cui et. al. introduced a content-based mesh based on the redundant wavelet[6]. However, it is non-trivial to retrieve the same content-based mesh generated in the encoder when decoding without high overhead, which makes deployment in the wavelet coder prohibitive. In our method, the first layer is always initialized with a regular mesh. There is high correlation between deformation of mesh layers. An efficient mesh topology coding strategy can be realised.

Secondly, an effective trade-off between motion coding and residual coding is of prime importance as indicated by the ’Crew’ sequence. The layered mesh provides efficient control of the trade-off. Futhermore, intensity control can be introduced using the existing mesh. Each grid point has an additional parameter for intensity scaling by which pixels inside the patch are interpolated.

Lastly, mesh-based coding is also an efficient model for very low bitrate coding with the advantages as mentioned. Adequate subsampling of each layer of mesh leading to a pyramid structure can provide additional improvement in bitrate for the coding of motion information.

6 Conclusion

We have described a simple yet effective algorithm that uses the frame difference generated from a mesh-based motion compensated image to identify regions of motion discontinuity. Motion estimation in these regions is refined using a finer mesh structure. It is notable that the proposed approach provides hierarchical refinement without additional overhead, and with no constraint on the movement of grid point positions. The algorithm can be combined with any regular mesh topology. This work shows an improvement over the single-layered mesh refinement technique.

References

1. Nakaya, Y., Harashima, H.: Motion compensation based on spatial transformations. IEEE Trans. CSVT. **4**(3) (1994) 339–356
2. Huang, C., Hsu, C.: A new motion compensation method of image sequence coding using hierarchical grid interpolation. IEEE Trans. CSVT. **4**(1) (1994) 42–51
3. Toklu, C., Erdem, A., Sezan, M., Tekalp, A.: Traking motion and intensity variations using hierarchical 2-d mesh modelling for synthetic object transfiguration. Graphical Models and Image Process. **58**(6) (1996) 553–573
4. Ohm, J., Schaar, M., Woods, J.: Interframe wavelet coding-motion picture representation for universal scalability. Signal Process.: Image Commun. **19**(9) (2004) 877–908
5. Secker, A., Taubman, D.: Highly scalable video compression using a lifting-based 3d wavelet transform with deformable mesh motion compensation. In: IEEE ICIP. Volume 3., Montreal Canada (2002) 749–752
6. Cui, S., Wang, Y., Fowler, J.: Mesh-based motion estimation and compensation in the wavelet domain using a redundant transform. In: IEEE ICIP. Volume 1., Rochester, New York (2002) 693–696

Very Low Bitrate Video: A Statistical Analysis in the DCT Domain

Mihai Mitrea[1,2], Françoise Prêteux[1], and Mihai Petrescu[1,2]

[1] ARTEMIS Department, GET/INT,
9, Rue Charles Fourier,
Evry 91011, France
{mihai.mitrea, francoise.preteux}@int-evry.fr,
mihaip_ro2@yahoo.co.uk
http://www-artemis.int-evry.fr
[2] Faculty of Electronics, Telecommunications and Information Technology,
POLITEHNICA University of Bucharest, Romania

Abstract. While generally the watermarking methods are designed to protect high quality video (*e.g.* DVD), a continuously increasing demand for protecting **v**ery **l**ow **b**itrate **v**ideo (VLBV) - *e.g.* in mobile networks, the video stream may be coded at 64kbit/s - is also met nowadays. In this respect, a special attention should be paid to the statistical behaviour of the video content. At our best knowledge, this paper presents the first statistical investigation on VLBV in the DCT (2D Discrete Cosine Transform) domain, thus identifying the particular way in which the VLBV obeys to the very popular Gaussian law. It also points to critical behaviour differences between the high and very low bitrate videos. These theoretical results are validated under the framework of watermarking experiments carried out in collaboration with the SFR mobile service provider (Vodafone group).

1 Introduction

While the 90's were characterised by an explosion of personal computer capabilities (processing power, storage capacity, printing quality and Internet connection speed), the last five years have testified a continuous expansion of the mobile phone paradigm. In fact, nowadays, a mobile phone is designed more like a computer than a traditional phone: the voice services are just a small part of the features now available, a mobile terminal allowing the user to have a video conference, to browse the Internet or to watch live television. Hence, music, video, and 3D characters are just some content examples that imposed themselves as a very important component of multimedia distribution to mobile terminals. As the copyright related issues are not alleviated in the mobile networks, sound watermarking methods should be designed to protect these content types. The present paper is part of a larger research study aiming at designing a watermarking method for VLBV. In this respect, to determine an accurate statistical model for such video is a first and very important step.

In its largest acceptation [1-3], *video watermarking* stands for the practice of imperceptibly altering a video sequence in order to embed a message. This embedded

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 99 – 106, 2006.
© Springer-Verlag Berlin Heidelberg 2006

message is referred to as *mark* or *watermark*. Generally, it conveys copyright information (*e.g.* the video owner, the number of allowed copies, the time when that video was sold) and should be generated starting from some secret information referred to as *key*. According to the targeted application, the size (in bits) of the copyright information may vary. When the embedded message does not alter the visual quality of the considered video, the watermarking procedure features *transparency*. The *robustness* refers to the ability of the watermark to survive signal processing operations. Two classes of such operations should be considered. The first class contains the common transformations applied to the video sequence, *e.g.* compression, change of file format, temporal cropping, colour reduction, *etc*. The second class is represented by the attacks. These are malicious transforms designed to make the watermark detection unsuccessful while preserving a good visual quality for the video. In this respect, StirMark [4] can be considered as the most harmful attack.

A special attention should be paid to the room in which the mark should be embedded. On the one hand, robustness requires the mark to be embedded into some video salient characteristics (*e.g.* low frequency DCT coefficients). On the other hand, transparency does not allow the salient characteristics to be drastically altered. Hence, to reach the trade-off between transparency and robustness is a core issue for any watermarking technique.

From the information theory point of view, a watermarking method is modelled as a noisy channel. The mark to be embedded is a sample from the input information source. The original video, the attacks and the mundane transformations applied to the marked video stand for a sample from the noise source. In order to design a good watermarking technique, the noise statistical behaviour should be known. For a large class of watermarking techniques [5, 6], the detection procedure is based on correlation (matched filters). This detection rule becomes optimal when the noise is Gaussian distributed. As the main noise component is represented by the original video itself, a previous study [7] determined the way in which the high bitrate video sequences can be modelled by ergodic Gaussian information sources and took advantage of this result in order to optimise a watermarking method [6].

At our best knowledge, this paper reports the first study aiming at modelling with the mathematical rigour the VLBV in the DCT domain. It reconsiders the method developed in [7] and determines whether and how the Gaussian law can be involved in such a statistical description.

The paper has the following structure. The statistical method is described in Section II. Section III presents the experimental results. Section IV brings into discussion the practical impact the VLBV modelling can have on watermarking and concludes the paper.

2 A Statistical Investigation in the DCT Domain

Maybe the most intensively & extensively used transform in image/video processing is the DCT: compression, indexing, retrieval, cryptography, watermarking are just some application fields in which it proved its opportunity. Therefore, finding an accurate statistical model for the DCT coefficients has always been a challenging matter [8-12]. All these studies regarded the values taken by some spatial frequencies corresponding to good quality still images. For high bitrate video, the values corresponding to the hierarchy of the DCT coefficients has also been investigated [7]. In the sequel, we shall investigate the DCT coefficient hierarchy, this time in order to bring information about the VLBV.

Be there the 2D random process which has as samples the frames in VLBV sequences. This random process will be investigated by defining R random variables, as follows. The DCT is individually applied to each frame and the coefficients thus obtained are sorted in a decreasing order of their values. The R random variables to be analysed are represented by the largest R values in such a hierarchy. These random variables are defined on the same probability space as the considered random process: each sample of the random process (*i.e.* each frame) leads to an experimental value for each random variable (*i.e.* a value for a rank).

The method is to be individually applied to each r, $r \in [1, R]$, random variable, *i.e.* to each hierarchy rank. Be there an L frame video sequence and be there an r rank, $r \in [1, R]$, arbitrarily chosen.

In order to overpass the dependency existing among successive frames in a video sequence, a periodical sampling is performed: when the D sampling period is large enough, a set of independent frames is obtained. By shifting the sampling origin, a partition into D classes, each of them with independent elements is obtained. These classes are *a priori* equally good in representing the video sequence. Hence, we shall first individually investigate the classes in the partition and then we shall fusion all these partial information. However, note that these classes are dependent.

For each of the D classes in the partition, we first verify whether its elements are Gaussian distributed or not, by means of a Chi-square statistical test [13]. As we have no *a priori* information about the mean value and the variance corresponding to the theoretical *pdf* (*p*robability *d*ensity *f*unction) to be tested, these two parameters should be estimated.

In the next step, we *a posteriori* check-up whether the considered D value is large enough so as to afford the independence among the elements in the same partition class. A Ro statistical test [13] is performed. As the Ro test should only be applied to Gaussian data, the result is meaningful only when the Chi-square test is passed.

The third step verifies the homogeneity of the elements in a class, from the variance point of view, by applying a Fisher F statistical test on equality between two variances [14]. When running the F test, the coefficients in a partition class are randomly split up into two sets. As an F test should be applied to independent Gaussian data, the result is meaningful only when the Chi-square and Ro tests are passed. The homogeneity of the elements in a class was also verified from the mean value point of view, by performing a Student T statistical test on equality between two means [14]. Here again, the elements of the class are randomly split up into two sets. As there is no *a priori* information about the theoretical variances, the T test should be applied to independent Gaussian data sets with equal variances (*i.e.* when the Chi-square, Ro, and F tests are passed).

When the above-mentioned tests are passed, it can be stated that the corresponding class does not refute the Gaussian behaviour for the considered r rank. However, there is no information whether the D Gaussian laws corresponding to the same rank would be identical or not. Hence, a homogeneity investigation among the D classes is considered. In this respect, we first verified the homogeneity of the D variances, by means of the F test. We ran $D-1$ such tests, each time on consecutive classes (*i.e.* first class *vs.* second class, second class *vs.* third class, and so on). Then, we resumed the experiment, by considering a T test, this time having in view the mean value homogeneity.

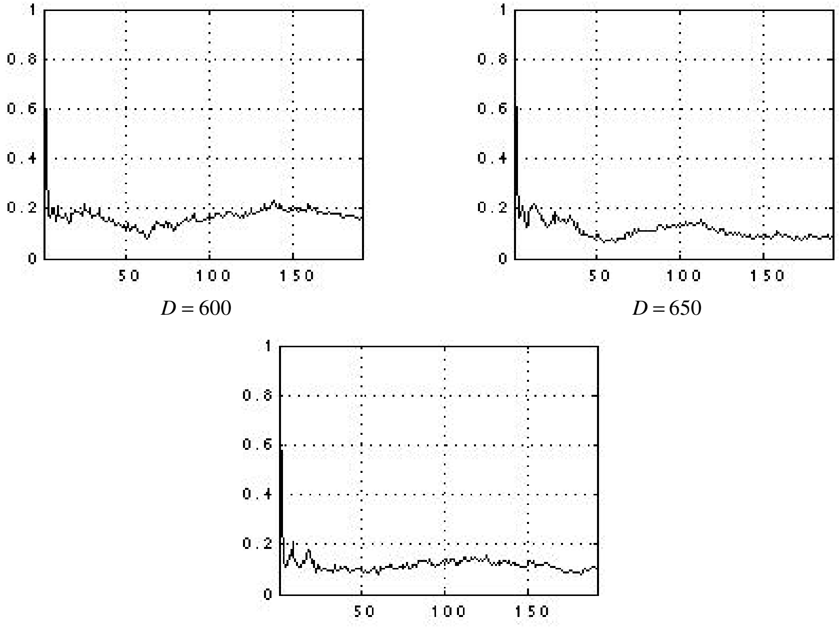
We can speak about a Gaussian law characterising the DCT decomposition only when all (or, at least, almost all) the above-mentioned tests are passed.

It can be claimed that the above presented procedure features mathematical rigour. In contrast to other types of statistical investigation [8-12], it makes no assumption either on pixel correlation or on the video sequence ergodicity. These two issues are here properly analysed, by means of the Ro tests and of the homogeneity investigation (the F and T tests).

3 Experimental Results

The experiments have been carried out on 10 video sequences of $L = 35000$ frames each (about 25 minutes each). They were DivX compressed, at 64kbits/s (the standard rate in telephony). Their frame width was 192 pixels (in concordance with the MOTOROLA V550 cell phone) and the aspect ratio was between 1.9 and 2.7. The ranks $r \in [1\,, R = 192]$ were under investigation.

As the sequences we considered are represented into the hue-saturation-value (*HSV*) colour space [15], the DCT is applied to the *V* component which is normalised to the $[0\,,1]$ interval.



$D = 600$

$D = 650$

$D = 700$

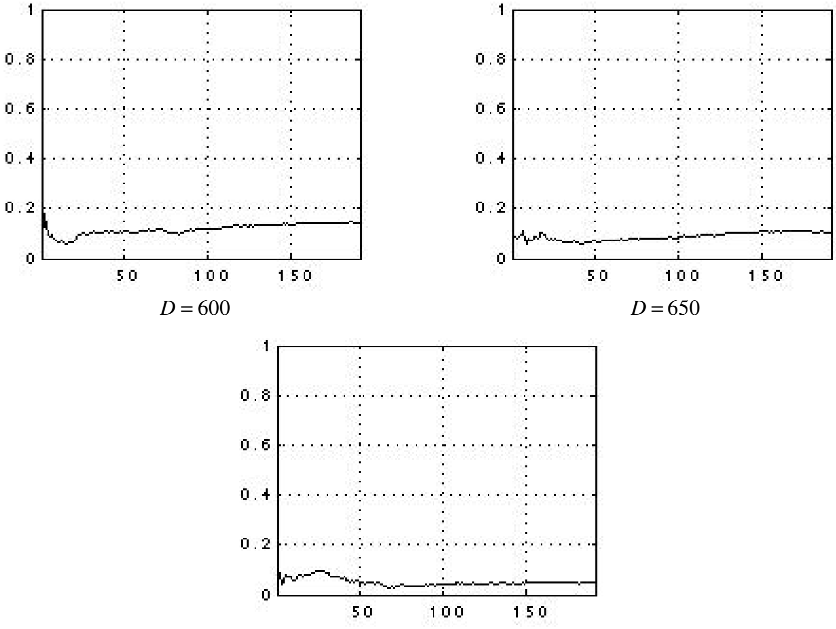
Fig. 1. The statistical investigation on VLBV: the abscissa represents the investigated rank and the ordinate stands for the relative number of Chi-square tests which were not passed

Fig. 1 presents the results obtained when applying the Chi-square tests to each class in the partition. The r rank, $r \in [1, 192]$ is represented on the abscissa. The three plots in Fig. 1 correspond to three values considered for the D sampling period, namely: $D = 600$, $D = 650$, and $D = 700$ frames respectively. For each rank, D Chi-square tests were applied. The ratio of the number of tests which are not passed to the D value (*i.e.* the relative number of tests which are not passed) is depicted on the y-axis. It can be seen that the Gaussian behaviour is not refuted for a large rank interval: when $r \in [10, 192]$ more than 80% of tests are passed.

The experiments synthesised in Fig. 2 were carried out in order to *a posteriori* validate the three above-mentioned numerical values for the D sampling period. The abscissa refers to the investigated rank $r \in [1, 192]$ while the ordinate stands for the relative number of Ro tests which were not passed (*i.e.* the ratio of the number of the tests which were not passed to the D value). When inspecting Figs. 2, it can be noticed the value $D = 600$ frames (*i.e.* 24s) is large enough so as to ensure the independence among the coefficients in a partition class.

The tests involved in the homogeneity investigation were successively passed: each and every time, more than 95% of tests were passed.

The results in Figs. 1&2 were obtained on a particular video sequence. Of course, their degree of generality is a crucial issue: we can not expect that the plots would be identical for all the video sequences but their overall behaviour should not depend on

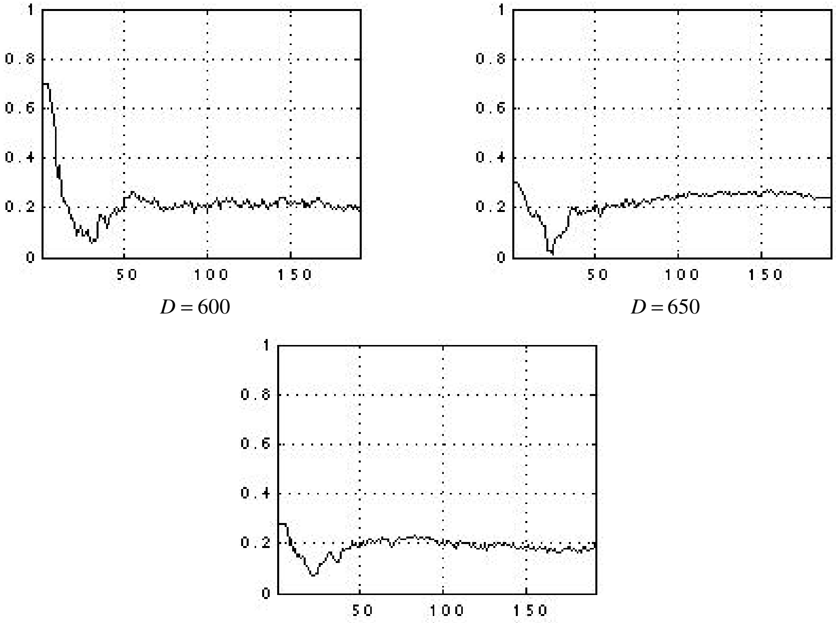


$D = 600$

$D = 650$

$D = 700$

Fig. 2. The statistical investigation on VLBV: the abscissa represents the investigated rank and the ordinate stands for the relative number of Ro tests which were not passed



$D = 600$

$D = 650$

$D = 700$

Fig. 3. The experiments in Fig. 1 are resumed for a different video sequence

the investigated video sequence. Fig. 3 depicts the results obtained when applying the Chi-square tests on a different video sequence. This time, for the ranks $r \in [50\,,192]$, a smaller number of tests were passed (about 75%). There were 3 video sequences (out of the 10) for which the $D = 600$ frame sampling period was too small: in order to obtain the independence among the investigated coefficients, a value $D = 650$ frames should be considered.

4 Final Remarks

This paper brings into evidence the fact that the Gaussian law can model a certain rank interval in the DCT hierarchy for the VLBV: when $r \in [10\,, 50]$, more than 80% of the Chi-square and Ro tests are passed. This percentage decreases to 75% (or even lower) when considering the coefficients with the $r \in [51\,,192]$ ranks. Hence, some zero memory Gaussian information sources approximating to VLBV are identified.

These results significantly differ from the ones obtained for high bitrate video [7]: in that case, the $15 < r < 120$ interval cannot support the Gaussian law, Fig. 4. Instead, the $r \in [120\,,192]$ rank interval features a very fine Gaussian behaviour. In [7] it was also stated that the non-Gaussian coefficients would feature a much larger redundancy than the Gaussian ones. The results now presented validate this suggestion: the VLBV has a much lower redundancy and a smaller number of non-Gaussian coefficients.

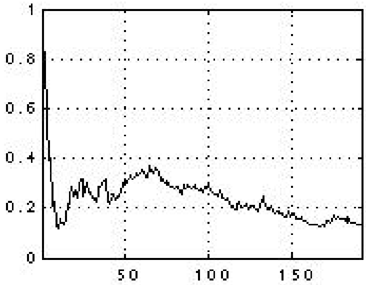


Fig. 4. The statistical investigation on high rate video

From the watermarking point of view, this statistical study can be a guide when selecting the right location to embed the mark at. For high rate video, the most suitable locations are $r \in [128\,,192]$ because: (1) they are Gaussian distributed, and (2) they correspond to ranks reaching the trade-off between transparency and robustness. When considering the VLBV, the finer Gaussian behaviour is featured by the $r \in [10\,,50]$ interval. A mark embedded at these locations would be more robust but it would also lead to stronger artefacts (*i.e.* artefacts that cannot be accepted for some applications). This statement is supported by two experiment types. Be there the video watermarking method in [6] and be there a high rate video sequence (*e.g.* 1024 kbit/s, 512×512 pixel frames). This sequence is compressed at 64 kbit/s, 192×80 pixel frames. In the first experiment, the mark was embedded in the $r \in [20\,,51]$ ranks, *i.e.* in non-Gaussian coefficients for high bitrate video and in Gaussian coefficients for VLBV. When detecting the mark after a StirMark attack, the VLBV video led to a 75% lower error rate. However, in both cases, transparency was very poor. In the second experiment, the mark was embedded into the coefficients with the ranks $r \in [128\,,159]$, *i.e.* into Gaussian coefficients for the high bitrate video and into coefficients which do not strictly obey the Gaussian law for VLBV. When imposing a transparency constraint, the VLBV could not withstand the StirMark attack. The high quality video behaved differently: it afforded both the StirMark robustness and a very fine transparency (no visual differences between the marked and the unmarked sequences).

To conclude with, the results in this paper may be also considered as proving the limitations some watermarking techniques [5-6] reach under the VLBV framework: the trade-off between transparency and robustness can no longer be identified. In order to overcome this problem, several solutions can be considered: (1) to change the detection rule; (2) to embed the mark in the DWT (Discrete Wavelet Transform) domain rather than in the DCT; (3) to consider an informed watermarking approach [16].

The application area of the VLBV statistical model here presented is not restricted to watermarking; for instance, it can also contain video compression, indexing & retrieval. The future work will focus on such applications, as well as on the VLBV investigation in the DWT domain.

Acknowledgement

This study is part of the TAMUSO (**TA**touage **M**ultimédia et ses **U**sages dans les ré**S**eaux m**O**biles) R&D contract between GET/INT and the SFR mobile service provider in France.

References

1. Cox I., Miller M., Bloom J.: Digital watermarking. Morgan Kaufmann Publishers (2002).
2. Katenbeisser S., Petitcolas F.: Information hiding – techniques for steganography and digital watermarking. Artech House (2000).
3. Arnold M., Schmucker M., Wolthusen S.: Techniques and applications of digital watermarking and content protection. Artech House (2003).
4. Petitcolas F., Anderson R., Kuhn M.: Attacks on copyright marking systems. David Aucsmith (ed.), Lecture Notes in Computer Science, Vol. 1525, Springer-Verlag, Berlin Heidelberg New York (1998).
5. Cox I., Kilian J., Leighton T., Shamoon T.: Secure spread spectrum watermarking for multimedia. IEEE Trans. on Image Processing, Vol. 6, No. 12 (1997) 1673-1687.
6. Mitrea M., Prêteux F., Vlad A.: Spread spectrum colour video watermarking in the DCT domain. JOAM (Journal of Optoelectronics and Advanced Materials), Vol. 7, No. 2 (2005) 1065-1072.
7. Mitrea M., Prêteux F., Vlad A., Fetita C.: The 2D-DCT coefficient statistical behaviour: a comparative analysis on different types of image sequences. JOAM, Vol. 6, No. 1 (2004) 95-102.
8. Pratt W.: Digital Image Processing. John Willey (1978).
9. Reininger R., Gibson J.: Distributions of the 2D-DCT coefficients for images. IEEE Trans. on Communications, Vol. COM–31, No. 6 (1983) 835-839.
10. Müller F.: Distribution shape of two–dimensional DCT coefficients of natural images, Electronics Letters, Vol. 29, No. 22 (1993) 1935-1936.
11. Joshi R., Fischer T.: Comparison of generalized Gaussian and Laplacian modeling in DCT image coding. IEEE Signal Processing Letters, Vol. 2, No. 5 (1995) 81-82.
12. Lam E., Goodman J.A.: A mathematical analysis of the DCT coefficient distributions for images. IEEE Trans. on Image Processing, Vol. 9, No. 10, (2000) 1661-1666.
13. Walpole R.E., Myres R.H.: Probability and statistics for engineers and scientists. MacMillan Publishing (1989).
14. Frieden B.R.: Probability, statistical optics and data testing. Springer-Verlag, New York (1983).
15. International Standard 15938-3 – Information Technology – Multimedia Description Interface, Part 3 Visual 2001. The MPEG-7 international standard, Geneva, Switzerland.
16. Miller M., Doerr G., Cox I.: Applying informed coding and embedding to design a robust high-capacity watermark. IEEE Trans. on Image Processing, Vol. 13, No. 6 (2004) 792-807.

H.264/AVC Based Video Coding Using Multiscale Recurrent Patterns: First Results

Nuno M.M. Rodrigues[1,2], Eduardo A.B. da Silva[3], Murilo B. de Carvalho[4], Sérgio M.M. de Faria[1,2], and Vitor M.M. da Silva[1,5]

[1] Instituto de Telecomunicações, Univ. of Coimbra - Pole II,
Coimbra 3030-290, Portugal
[2] ESTG, Inst. Politécnico Leiria, Apartado 4163, Leiria 2411-901, Portugal
[3] PEE/COPPE/DEL/Poli, Univ. Fed. Rio de Janeiro,
Caixa Postal 68504 - Cidade Universitária, Rio De Janeiro 21945-970, Brazil
[4] TET/CTC, Univ. Fed. Fluminense, Campus da Praia Vermelha,
Rua Passo da Pátria, 156, São Domingos, Niterói 24210-240, Brazil
[5] Dep. of Electrical and Computer Engineering, Univ. of Coimbra - Pole II,
Coimbra 3030-290, Portugal
nuno.rodrigues@co.it.pt, eduardo@lps.ufrj.br, murilo@telecom.uff.br,
sergio.faria@co.it.pt, vitor.silva@co.it.pt

Abstract. The Multidimensional Multiscale Parser (MMP) algorithm has been proposed recently as a universal data coding method. MMP has proved to be a very powerful coding method for images, as for other types of signals. Experimental tests showed that MMP is able to achieve better results than the traditional transform-based image coding methods, particularly for images that do not have a low-pass nature.

These promising results motivated the use of MMP for residual error encoding in hybrid video coding algorithms. This paper presents the first results of these experiments, performed using a H.264/AVC based video encoder, but using MMP to encode the motion compensated residual data, for the P and B slices.

Experimental results show that, even in this not fully optimised version, this method is able to achieve an approximately equivalent performance to the H.264/AVC. This demonstrates that MMP is an alternative to the transform-quantisation paradigm for hybrid video coding that is worth investigating.

1 Introduction

Hybrid video coding schemes have been almost ubiquitous in video coding standards. They mostly use block based motion estimation and compensation to reduce the energy of the residual image. The error image is then encoded using transform coding and quantisation. Such residual error encoding is a legacy from the top image encoding methods, which traditionally use algorithms based on the transform-quantisation paradigm with excellent results.

The most recent standard for video coding, H.264/AVC (H.264) [1] also uses a transform based residual encoding method, that has been highly optimised for coding efficiency.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 107–114, 2006.
© Springer-Verlag Berlin Heidelberg 2006

In this work we introduce an algorithm to encode the motion predicted data in an H.264 based video coder, which is based on an alternative paradigm to the transform-quantisation. This algorithm is referred to as Multidimensional Multiscale Parser (MMP) [2], because it uses an adaptive dictionary to approximate variable-length input vectors. These vectors result from recursively parsing an original input block of the image. Scaling transformations are used to resize each dictionary element to the dimension of the block segment that is being considered.

Previous results [2] show that MMP performs well for a wide variety of input images, ranging from smooth grayscale images to text and graphics. This lends it a universal flavour. Therefore, one expects that it should also perform well for encoding residual images. This was confirmed by results in [3], where MMP is used to encode intra prediction residuals. This motivated the use of MMP for encoding the motion compensated residual data in the H.264 video coder, replacing the adaptive block size transform (ABS) defined in this standard. We refer to it as the MMP video encoder.

This paper presents the first results of this encoder. MMP is used to encode the motion compensated residual image in P and B slices. Since our aim is to assess the performance of MMP for motion compensated residual data, the intra macroblock (MB) residues are encoded using the original H.264 transform. All other syntax elements are also encoded using the techniques defined in H.264 (JM9.3) reference software [4].

Results of the first tests have shown that the MMP video encoder has an overall performance comparable to that of H.264. Taking into consideration that in these tests the rate-distortion decisions are the same ones used in H.264, we believe they have not been optimised for the use of MMP. This suggests that there might be some room for improvement, and therefore that it is worth investigating MMP as an alternative to the present transform-quantisation paradigm in video coding.

In the next section we briefly present the MMP algorithm for image coding. Section 3 describes the new MMP video encoder, and in section 4 the first experimental results are presented and compared with the ones of H.264 high profile. Conclusions of this work can be found in section 5 along with a brief outline of planned future work.

2 The MMP Algorithm

Although the MMP algorithm was initially proposed as a generic lossy data compression method, it is easily expandable to work with n-dimensional data, and has been successfully applied to two dimensional data. In this section we describe the most important aspects of the MMP algorithm applied to image coding. An exhaustive description of the method can be found in [2].

MMP is based on approximations of data segments (in this case image blocks), using words of an adaptive dictionary $\mathcal{D}$ at different scales. For each block X^l in the image, the algorithm first searches the dictionary for the element S_i^l that minimises the Lagrangian cost function of the approximation. The superscript l means that the block X^l belongs to *level* l of the segmentation tree (with dimensions $(2^{\lfloor\frac{l+1}{2}\rfloor} \times 2^{\lfloor\frac{l}{2}\rfloor})$). Square blocks, corresponding to even levels, are segmented into two vertical rectangles.

The algorithm then segments the original block into two blocks, X_1^{l-1} and X_2^{l-1}, with half the pixels of the original block, and searches the dictionary of level $(l-1)$ for the elements $S_{i_1}^{l-1}$ and $S_{i_2}^{l-1}$ that minimise the cost functions for each of the sub-blocks.

After evaluating the rate-distortion (RD) results of each of the previous steps, the algorithm decides whether to segment the original block or not. Each non-segmented block is approximated by one word of the dictionary (S_i^l). If a block is segmented, then the same procedure applied to the original block is recursively applied to each segment.

The resulting binary segmentation tree is encoded using two binary flags: flag '0' represents the tree nodes, or block segmentations and flag '1' represents the tree leafs (sub-blocks that are not segmented). These flags are not used for blocks of level 0, that can't be further segmented.

The binary tree is encoded using a preorder approach: for each node, the sub-tree that corresponds to the left branch is first encoded, followed by the right branch sub-tree. In the final bit-stream, each leaf flag is followed by an index, that identifies the word of the dictionary that should be used to approximate the corresponding sub-block. These items are encoded using an adaptive arithmetic encoder.

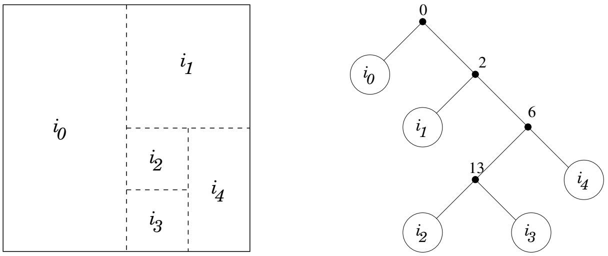


Fig. 1. Segmentation of a 4×4 block and the corresponding binary tree

Figure 1 represents the segmentation of an example block and the segmentation tree that MMP uses to encode it. In this example, $i_0 \ldots i_4$ are the indexes that were chosen to encode each of the sub-blocks, and so this block would be encoded using the following string of symbols:

$$0 \quad 1 \quad i_0 \quad 0 \quad 1 \quad i_1 \quad 0 \quad 0 \quad i_2 \quad i_3 \quad 1 \quad i_4 .$$

The RD optimisation of the segmentation tree, $\mathcal{T}$, that is used to encode each block, is performed evaluating the Lagrangian cost for every segmentation decision, given by $J(\mathcal{T}) = D(\mathcal{T}) + \lambda R(\mathcal{T})$, where $D(\mathcal{T})$ is the distortion obtained when using $\mathcal{T}$ and $R(\mathcal{T})$ is the corresponding rate.

Unlike conventional vector quantisation (VQ) algorithms, MMP uses *approximate block matching with scales* and an *adaptive dictionary*.

Block matching with scales is an extension of the ordinary pattern matching, in the sense that it allows the matching of vectors of different lengths. In order to do this, MMP uses a separable scale transformation T_N^M to adjust the vectors' size before trying to match them. For example, in order to approximate an original block X^l using one block S^k of the dictionary, MMP has to first determine $S^l = T_k^l[S]$. Detailed information about the use of scale transformations in MMP is presented in [2].

MMP uses an adaptive dictionary that is updated while the data is encoded. Every time a block is approximated by the concatenation of two dictionary blocks, of any given level, the resulting block is used to update the dictionary, becoming available to encode future blocks of the image, independently of their size. This updating procedure for the dictionary uses only information that can be inferred by the decoder exclusively from the encoded segmentation flags and dictionary indexes. Thus, MMP has the ability to learn the patterns that previously occurred on the image, adapting itself to the data being encoded. This characteristic gives it a universal flavour.

3 The MMP Based Video Encoder

In this section we describe the main features of the MMP video encoder. It is based on the JM9.3 reference software of H.264 video coding standard [4]. All the encoding modes are inherited from H.264, as well as the rate-distortion (RD) encoding decisions.

The main difference between the MMP video encoder and H.264 is related to the motion compensated residue data encoding method for the P and B macroblocks: MMP replaces the original DCT integer transform defined in [1].

One important feature of H.264 is the use of adaptable block size transforms. Such transforms provide a significant gain in coding efficiency when compared with the fixed size blocks (either 4×4 or 8×8) used by its predecessors. Such scale adaptability allows saving bits by the use of large blocks where the residual is mostly uniform, while providing good coding accuracy through the use of small blocks in the cases where the residue data is more detailed. It is important to note that scale adaptability is a feature inherent to MMP. In fact, this is one of the strong reasons for its good coding performance.

3.1 Intra Encoding

The H.264 recommendation defines three different partition block sizes for intra MB's: 16×16, 8×8 and 4×4. Intra prediction is done using four possible prediction modes for $Intra\_16 \times 16$ macroblocks and nine prediction modes for the other two partition sizes.

Previous tests compared the efficiency of MMP and H.264 in encoding the intra residue for still digital image coding. The MMP-Intra method uses a set of prediction schemes similar to those defined for H.264 intra encoding, but encodes the prediction residue with MMP. Details about this method and experimental results comparing its performance against H.264 and JPEG2000 are presented in [3].

As previously stated in the Introduction, in this version of the MMP video encoder all the intra MB's are encoded using exactly the same procedure as H.264, including the same integer DCT transform. The reason for this is that we are mainly interested in assessing the performance of MMP for motion compensated residual coding. We do so by comparing the coding efficiency of MMP video with the one of the H.264 encoder. This comparison is only effective if the reference frames used by both methods are the same. This would not be the case if MMP was used for intra MB coding.

3.2 Inter Encoding

Inter MB coding involves two major steps: motion estimation and coding of the motion compensated residue. Motion estimation consists in the search of the motion vector that allows the best result for the block residue, in a RD sense. H264/AVC uses quarter-pixel precision for the motion vectors and performs the motion estimation in three separate steps. First a full search determines the best motion vector with full pixel precision. After this, the best motion vector with half and quarter pixel precision is determined around the position obtained in the previous step.

In addition, motion compensation in H.264 is done using one of seven modes, that are related to the partitioning possibilities of a 16x16 luma macroblock. Each partition of a MB has its own motion vector, that is used in the motion compensation of the corresponding sub-block. Thus, motion estimation in H.264 implies the optimisation not only of the motion vectors, but also of the partition block size that is used for motion compensation.

The existence of several partition modes allows the motion compensation to be performed using a block size that optimises the distortion of the motion predicted decoded residue versus the bit-rate coding cost, corresponding to the motion compensation data plus the residue transform coefficients. In H.264, the cost of each mode is estimated by evaluating the distortion of the transform coding of the residues, either by using the sum of absolute differences (SAD) or the sum of absolute transformed differences (SATD).

The current version of the MMP video encoder uses exactly the same motion estimation procedure. This has implications on the efficiency of the MMP encoder, because the motion estimation process returns a set of motion vectors that are the best in the "DCT point of view", meaning that these vectors minimise a cost function where the distortion is determined using the SA(T)D and the rate takes into account the cost of the DCT coefficients. These measurements are not related to those produced by the MMP coding, and we can expect a performance loss due to this fact. Nevertheless, this process favours a direct comparison between the encoding efficiency of MMP and the DCT, because the residue patterns that are generated by the motion estimation/compensation process tend to be approximately the same for both encoders. One should bear in mind, though, that we can expect an improvement in performance once the motion estimation process uses a cost function more related to the MMP characteristics.

H.264 encodes the motion compensated residue with a block size for the transform coding that depends on the partition size that was used for the motion estimation. The MMP video coder does not take this partition size into account and considers the entire 16×16 residue block, that is composed by the concatenation of the several partition blocks. This is not a problem because MMP is able to segment the original 16×16 block in a way that optimises the distortion of the decoded residue. Once again, the adaptability of the MMP encoding algorithm plays an important role in the efficiency of the method.

All motion compensation information, like the partition modes and the motion vectors for each block, is transmitted by the MMP video encoder using the same techniques as H.264, as described in [1].

4 Experimental Results

The MMP video encoder described in the previous section was implemented and experimental tests were performed.

In this test, MMP video coder uses six independent dictionaries: one for each of the YUV components of the P and B slices. The use of different dictionaries allows each of them to efficiently adapt to the specific predicted error block patterns of each type of source data. This has the additional advantage of limiting the size of each dictionary, reducing the computational complexity.

When used to encode prediction error blocks, MMP uses initial dictionaries in the scale 1×1 (level 0) with values in the range from -255 to 255. The initial dictionaries for the other levels are obtained from this one by scale transformation. The scale transformation and dictionary updating procedure are the same as those described in [2].

The MMP coder was compared with the H.264 high profile video coder using version 9.3 of the reference software. Both encoders were tested using the first 99 frames of the CIF Foreman sequence, 4:2:0. Only one I frame was used with one skipped frame and one B frame (I B P B P pattern). We used the variable bit rate mode, testing the encoders for several quality levels of the reconstructed video sequence. This was done by varying the QP parameter for the I/P and B slices.

Figure 2 represents the average PSNR for the luminance component of P and B frames versus the average number of bits used to encode each frame. For the P frames the RD curves are very close, indicating that, even using sub-optimal rate-distortion decisions, MMP can perform as well as the ABS transform. Figure 2 also shows that the MMP video coder tends to achieve better RD results for coding B frames than H.264. This can be explained by the fact that B frames use bidirectional motion estimation, which generates low energy residue patterns. The H.264 encoder uses a coarser quantisation scheme for these patterns. On the other hand, MMP is able to “learn” these residue patterns very efficiently and use them along the sequence, tending to encode such residue blocks better than H.264.

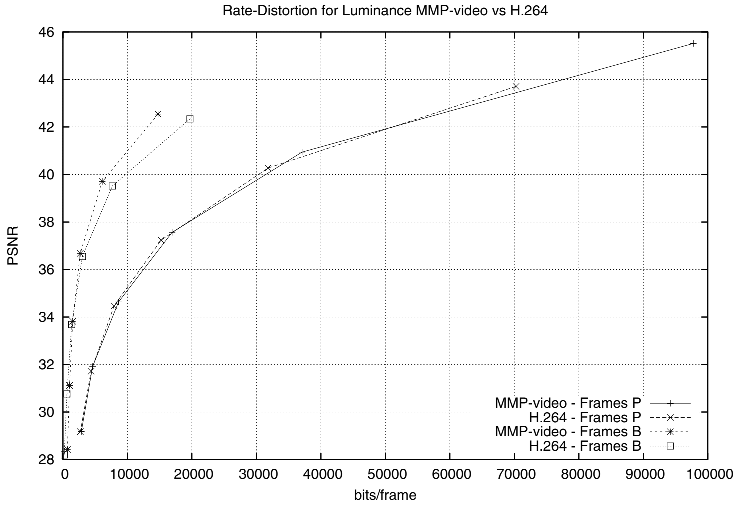


Fig. 2. Average luminance PSNR versus average number of bits per frame for P and B frames

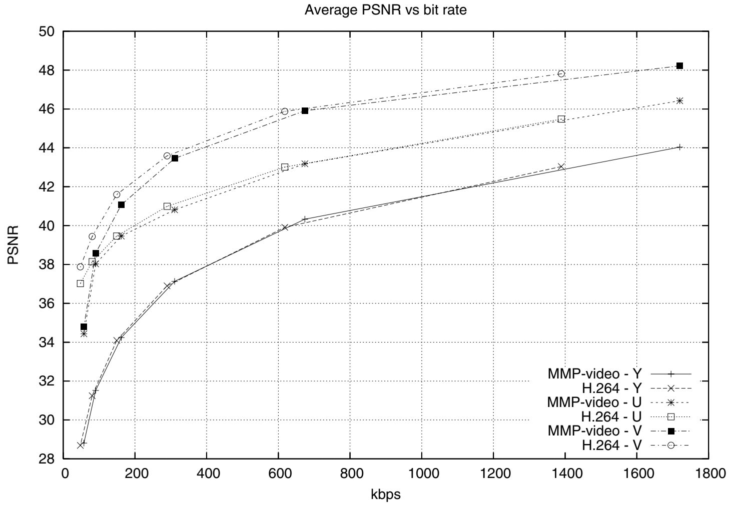


Fig. 3. Average PSNR versus bit rate for the YUV frame components

Figure 3 plots the average RD curves for each of the colour components. We can see that for the luminance component, both encoders have a similar performance, but there is some loss for the chroma components for the MMP video coder. An explanation for this is the fact that, since the chroma components are much smaller and have much less energy than the luminance, then the chroma dictionaries are not able to adapt to the residue patterns, specially at low rates. However, for high bit rates, MMP segments the block and compensates the small-size dictionaries.

A possible way to overcome this problem would be the use of a single dictionary for both chroma components, or even to devise ways of making the dictionaries to grow faster (for example, by introducing extra blocks related to the original pattern).

5 Conclusions and Future Work

This paper presents the first results of the use of the Multidimensional Multiscale Parser (MMP) for hybrid video coding. MMP is used instead of the integer DCT in a H.264/AVC based video coder, to encode the prediction residual data of the motion compensated macroblocks.

Experimental results have shown that the MMP video coder is able to achieve better results than H.264 for the luminance component of B and P slices, but still has some losses for the chroma components. Nevertheless, the current results demonstrate that MMP is an alternative to the integer DCT used by H.264 that is worth investigating.

The results presented in this paper show much room for further improvements, because the MMP video encoder has not yet been thoroughly optimised in a rate-distortion sense. Future work will address this issue, as well as several other questions that are relevant to the performance of the MMP video coder. Among them, one can distinguish the use of adaptive block size MMP to encode the motion compensated residue partition blocks, the use of MMP on intra MB's and the optimisation of the deblocking filter for the MMP-video reconstructed frames.

References

1. Draft of Version 4 of H.264/AVC ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 part 10) Advanced Video Coding). Document JVT-N050d1. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6) (2005)
2. de Carvalho, M., da Silva, E. Finamore, W.: Multidimensional Signal Compression using Multiscale Recurrent Patterns. Elsevier Signal Processing, (82), (2002), 1559-1580
3. Rodrigues, N., da Silva, E., de Carvalho, M., Faria, S., Silva, V., Universal Image Coding using Multiscale Recurrent Patterns and Prediction. IEEE International Conference on Image Processing, Genova, Italy, (2005), vol. 2, 245-248
4. *http://iphome.hhi.de/suehring/tml/download/*

Video Streaming in Electron Microscopy Applications

Francesca Mighela, Cristian Perra, and Massimo Vanzi

DIEE, Department of Electrical and Electronic Engineering,
University of Cagliari,
Piazza D'Armi, Cagliari 09123, Italy
{f.mighela, cperra, vanzi}@diee.unica.it

Abstract. This paper will explain the remoting process of a Scanning Electron Microscope (SEM). Two particular topics will be highlighted: the development of the video streaming architecture; the attempt to give the start for a new standard microscopy: the remote microscopy. Experiments have been performed on a commercial SEM platform.

1 Introduction

The project takes origin from two considerations: the importance of electron microscopy and the difficult to perform electron microscopy. In fact, it is one of the most useful techniques for many science field applications. Some examples are: microelectronics (in particular, reliability of microelectronic components), geology, mineralogy, physics, chemistry, and biology, also thank to the last instruments generation, able to work at partial environmental pressure. However, not each potential user could be able to buy and manage a Scanning Electron Microscope, as well as the necessary knowledge to perform the best analysis requires a lot of time to be reached.

In order to surpass these limitations, some microscopists have started to think to the remote microscopy [1], [2], [3] and [4].

This paper starts with a brief introduction about the reasons of the remote microscopy, and the problems concerning the research topics, like the access microscope and the public network capability. It continues with the principal subject: the video transmission architecture. Thus, there is a descriptive part of the video hardware of the instrument, and then a description of planned architecture for this research project. The architecture, to satisfy the characteristics required for an efficient and functional remote control, the real time transmission and the full resolution in relation to the commercial network connection available at the moment, has been identified into a dedicated transmission video streaming technology.

2 The Remote Microscopy

The entire project is based on the real necessity of remote microscopy; the different reasons are:

- To give the possibility to use the instrument remotely;
- To improve the collaboration between university research centres and the companies;

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 115 – 120, 2006.
© Springer-Verlag Berlin Heidelberg 2006

- To satisfy the university courses requirements, where the students number increases every year, making impossible to perform the lessons directly into the laboratory rooms.

Nevertheless, the tele-microscopy technology is still not diffused into the electron microscope world; the absence of any standard for SEM-based tele-microscopy also implies some strategic choices on the accessibility level that should be provided, where a strongly dedicated architecture, or a completely open one, define its boundaries.

A considerable limitation of the diffusion of remote microscopy is that the bandwidth required to transmit SEM data is about 100Mbps, a value that has made, in the past, the practical barrier to effective full-resolution, full-real-time data transmission. Nowadays, that limit has being surpassed by the band transmission speed currently available into local network (like University LAN), but not yet into public network.

3 The Project

The first step of the research program has been the choice of the instrument characteristics.

The instrument meets some particular requirements, like full digital control and large diffusion and helpfulness. It is straightforward to see as any SEM of the most recent generations is the perfect answer.

The second step has been the choice of the remote application and its software implementation.

Last but not least, the choice of the video transmission architecture; its implementation is not yet started.

3.1 The Instrument

The chosen platform is an instrument completely driven via software, and making possible to realize a remote sensing for each instrument functionality.

Figure 1 shows the instrument block diagram, with the internal network connections between each block.

3.2 The Remote Application

The microscope remoting project consists of implementing a server-client application at real time and full resolution.

The server application should reside on a PC physically close to the instrument that will work as a router between the microscope intranet and Internet. This PC is connected through the HUB 2 to the Support Network (see figure 1).

Remote control increases the microscope security and the possibility to have a major control of the remote users' actions. In fact, the server application will have a special tool to enable a client connection, and each client would logon with a private user name and password.

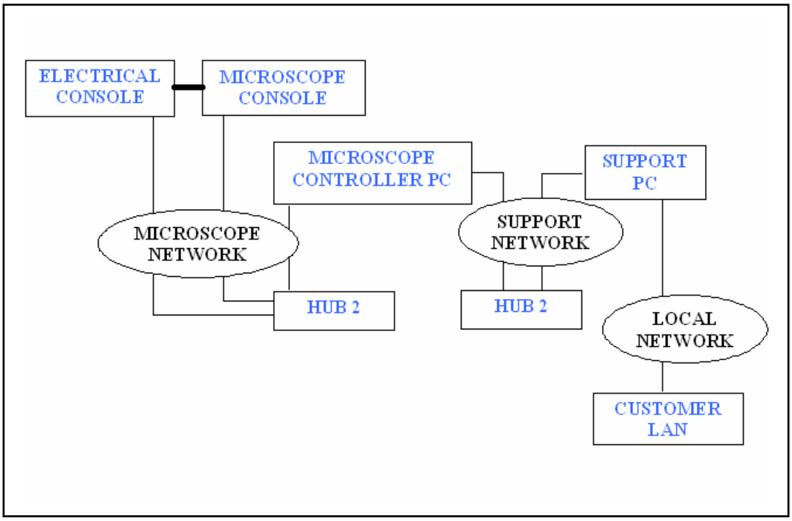
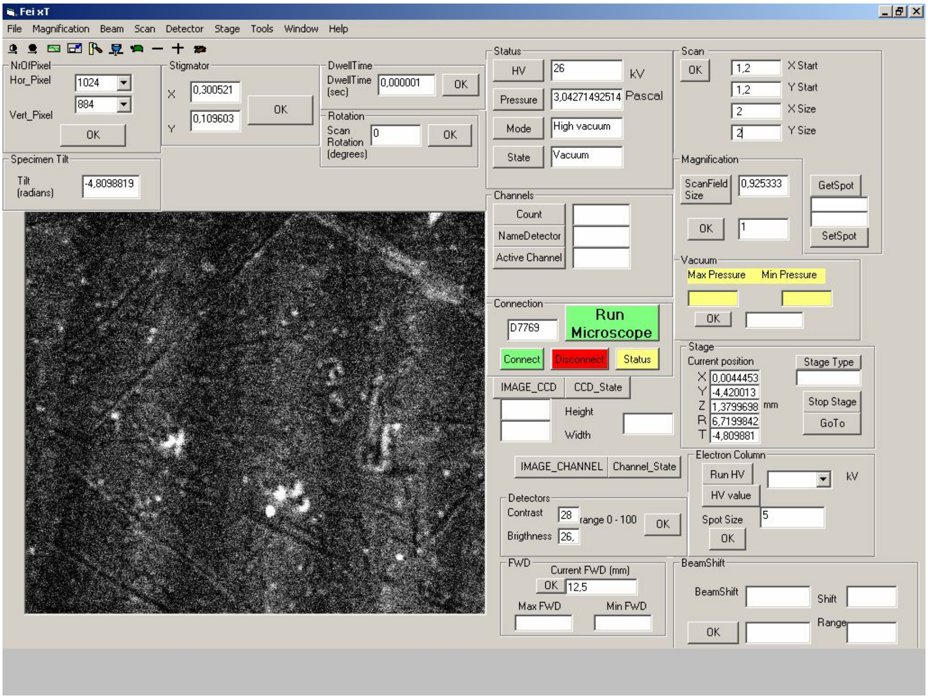


Fig. 1. Instrument Block Diagram



NrOfPixel
Hor_Pixel 1024
Vert_Pixel 884

Stigmator
X 0,300521
Y 0,109603

DwellTime
DwellTime (sec) 0,000001

Rotation
Scan Rotation (degrees) 0

Specimen Tilt
Tilt (radians) -4,8098819

Status
HV 26 kV
Pressure 3,04271492514 Pascal
Mode High vacuum
State Vacuum

Scan
X Start 1,2
Y Start 1,2
X Size 2
Y Size 2

Magnification
ScanField Size 0,925333
1

Vacuum

Connection
D7769

Stage
Current position
X 0,0044453
Y -4,420013
Z 1,3799698 mm
R 6,7199842
T -4,809881

Electron Column
kV
Spot Size 5

Detectors
Contrast 28 range 0 - 100
Brigthness 26,

FWD
Current FWD (mm) 12,5

Fig. 2. Client Application User Interface

Figure 2 sketches the client application User Interface.

The client application UI is divided into two main parts: the first part communicates with the commands instrument that is represented by the buttons, the dialog boxes, etc. Each of these commands corresponds to an instrument command in the original UI and communicates with it using the special library provided by the Manufacturer. The role of second part is to display the video coming out to the graphics system of Microscope Controller PC.

3.3 The Video Transmission

In the chosen platform the image displayed on the Microscope Controller PC is created in the overlay memory of the graphics card: the data are copied there from the frame-grabber image processor. The image on the screen is then result of the normal image generated by graphics card (user interface) and mixed-in overlay image from frame-grabber card.

This overlay operation inhibits the separated transmission of the two different generated images.

This configuration could be useful for making available a separated video output. This video should be managed to be transmitted to the PC server. The need of managing is due to the network capability. Also if the own microscope network is an Ethernet 100Mbps network, the public network is not yet able to guarantee this transfer data speed.

The video management starts at the frame grabber output, since the video coming into the frame grabber is a non standard video; then it is necessary to transform the output into a standard video to be read from the graphics card (e.g. standard SVGA). The frame grabber Manufacturer gives with it some special library to extract the video from the frame grabber; using these it is possible not only extract the digitized video, but also define the acquisition parameters and acquire images.

The parameters are identified, in example, into the number of frame buffers and into the image size. It is also possible to select the image file format between several choices.

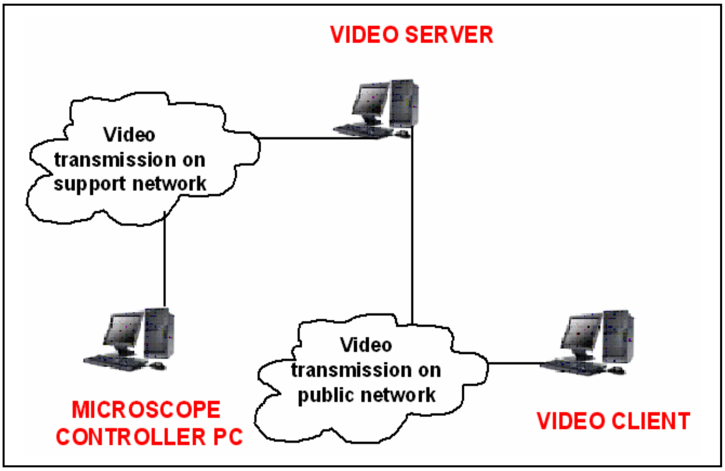

VIDEO SERVER
Video transmission on support network
MICROSCOPE CONTROLLER PC
Video transmission on public network
VIDEO CLIENT

Fig. 3. Video Transmission Architecture

This video extraction is performed from the instrument software to display the video on the Microscope Controller PC monitor. The intention is to replicate this extraction for sending the video to another application that will forward it to the PC remote server. At this point, the server application takes the video, and the client application will de able to manage the instrument and to see the video in real-time when connected to it. To do this, the video transmission architecture shown in Figure 3 has been designed. This is a dedicated architecture.

There is a video streaming server resident on the PC Remote Server that acquires the image from the Microscope Controller PC. The server streaming sends the video on the network and a nearly real time video is available for the remote client. On the remote user, a client streaming will be activated. The most recent video coding standard will be used as video streaming encoder: the H.264/AVC [5], [6].

3.4 Considerations

The video coming into the frame grabber is in slow scan mode; the video output is a digitized video, in example a SVGA standard video.

This video has several characteristics; in fact, to respect the microscope performance, it is possible to select different values of resolution and frame rate. The possible resolutions are: 512x442, 1024x884, 2048x1768, and 3584x3094 pixels; the frame period could be set from the minimum value, 0.1 seconds, until the maximum value, 946.74 seconds.

These values are necessary to determine the encoder compression setting for the transmission. Several tests carried out on the microscope have shown that the network average bandwidth is about 10Mbps.

Assuming that the microscope generates 10 frames per second, the better transmission condition is with a resolution of 512 x442 pixels, vice versa the worst case is a resolution of 3584 x3094 pixels.

Performing some simple calculations, in the first case the resultant compression ratio is 1:1.81, in the second it is1:88.

These are some calculations and considerations about the microscope performances and the network performances. Normally, the microscope works at 1024 x884 pixels resolution and 10 frames per second, because this reaches a good compromise between noise reduction and scan speed. In this situation, the compression ratio is 1:7.2, which assures that the remote user will receive the video with an excellent quality.

3.5 A Standard as Objective

The remote microscopy as it is explained in this paper is not yet a commercial product. The only type of remote application has been developed to remove the operator from the instrument. In fact, depending on the kind of instrument and on the kind of analysis, the presence of operator in front of the microscope could create some negative interference, or can be dangerous for the operator himself (e.g. laboratory classified as P3 and P4). Then, it should be useful to remove the operator, but it is necessary just to place him into a room near to that of the instrument.

This work aims to open the way to a new concept of remote microscopy giving the basic ideas to realize a tool of applications that should be a standard option for each electron microscope, both scanning and transmission.

4 Results

The results are very important in order to confirm the possibility to respect the features required for an efficient remoting microscope. The most critical phase, the video transmission, has been easily resolved thank to the new standard encoder H.264/AVC. In fact, the compression ratios obtained for each resolution guarantee always a clear video transmission. Obviously, in the max resolution case, the video will lose definition and resolution, but anyway it will be comprehensible; in the other cases, the resolution loss is not significant for the quality of the video displayed by the client.

5 Conclusions and Perspectives

The research project shows that the remote microscopy is useful and easy to achieve. Moreover, it confirms the relevant role that the remote microscopy could acquire, especially in relation to a recent filed remote control application: the collaborative environment through computing [7]. This is a solution implemented to overcome geographical distances between collaborating laboratories and to facilitate data sharing. However it is a relative recent argument, the Scientific International Community has just played much attention on it, developing several collaborative environments enjoying European Community Founds.

In order to integrate this remote control with the collaborative environments, the perspectives are: to terminate the server/client application; to develop the video streaming architecture; to manage the first result to achieve a plug and play application for a collaborative environment with the commercial SEM used for this project. The increasing of network capability of internet connection accessible in the microscope laboratory is a fundamental point to have a better remote microscopy.

References

1. http://www.itg.uiuc.edu/publications/techreports/03- 002/03-002.pdf
2. http://www.itg.uiuc.edu/publications/techreports/03 -001/03-001.pdf
3. http://bugscope.beckman.uiuc.edu
4. Van Balen, A., et al. : A Collaborative Environment for Users of Advanced Scientific Instruments. Third International Conference on Creating, Connecting and Collaborating through Computing, (C5 '05), (2005) 83-90
5. 5. Sullivan, G.J., Topiwala, P., Luthra, A.: The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extentions. SPIE Conference on Applications of Digital Image Processing XXVII Proceedings (2004)
6. Wiegand, T., Sullivan, G. J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC Video Coding Standard. IEEE Transaction on Circuits and Systems for Video Technology Proceedings (2003) 560-576
7. Van Balen, A., Morgan, C.L.: Details from a Distance: Seeing Angstroms at 10,000 Kilometers. INET Conference Proceedings (2000) 386-397

A New Generic Texture Synthesis Approach for Enhanced H.264/MPEG4-AVC Video Coding

Patrick Ndjiki-Nya, Christoph Stüber, and Thomas Wiegand

Image Communication Group, Image Processing Department,
Fraunhofer Heinrich-Hertz-Institut,
Einsteinufer 37, Berlin 10587, Germany
{ndjiki, stueber, wiegand}@hhi.de

Abstract. A new generic texture synthesis approach, which is inspired by the work of Kwatra et al. [1], is presented in this paper. The approach is hierarchical, non-parametric, patch-based and for that applicable to a large class of spatio-temporal textures with and without local motion activity. In this work, it is shown, how the new texture synthesizer can be integrated into a content-based video coding framework. That is, the decoder reconstructs textures like water, grass, etc. that are usually very costly to encode. For that, the new texture synthesizer in conjunction with side information that is generated by the encoder is required. The reconstruction of above-mentioned textures at the decoder side basically corresponds to stuffing holes in a video sequence. Spurious edges are thereby avoided by using graph cuts to generate irregular contours at transitions between natural and synthetic textures and preferably place them (the contours) in high-frequency regions, where they are less visible.

1 Introduction

Content-based video coding approaches typically partition a sequence into coherent regions given spatial, temporal, or spatio-temporal homogeneity constraints. For that, a texture analysis module is required at the encoder. Homogeneous segments are typically described using compact attributes like color, texture or motion features, which are transmitted to the decoder as side information. The decoder calls the synthesis counterpart of the texture analyzer to regenerate aforesaid regions using the side information. Significant bit rate savings can be achieved with this approach, while preserving high visual quality of decoded data, as is shown in [2]. This paper focuses on the texture synthesis module, assuming the texture analysis issue to be solved. Texture synthesis approaches can be divided into two categories: parametric and non-parametric. In both synthesis categories, the pdf of given texture examples is approximated and sampled to generate new, perceptually similar texture samples. The example textures are thereby assumed to be large enough to capture the statistics of the underlying infinite texture. Parametric synthesis approaches approximate the pdf using a compact model with a fixed parameter set. Such approaches entail helpful hints w.r.t. identification and recognition scenarios. Non-parametric synthesis approaches do not explicitly model the pdf of the texture examples, they rather directly match the neighborhood properties of the given sample or patch to synthesize

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 121 – 128, 2006.
© Springer-Verlag Berlin Heidelberg 2006

with the example texture. One of the selected candidates, with perceptually similar neighborhood properties to the sample or patch to synthesize, is then chosen for synthesis (pdf sampling). Not only do non-parametric synthesis approaches yield better synthesis results than parametric algorithms, also can they be successfully applied to a much larger variety of textures [1]. In terms of compactness of texture representation, non-parametric synthesis algorithms are typically less efficient than parametric ones. Thus, in the content-based video coding framework, parametric synthesis methods are usually used [3]. We will show, however, that non-parametric synthesizers can be used at the decoder to achieve good video quality.

The remainder of the paper is organized as follows. In Sec. 2, the hierarchical texture synthesis approach will be presented in-depth. The integration of the synthesizer into an H.264/MPEG4-AVC [4] video encoder, w.r.t. the GOP (Group of Pictures) structure and the side information generation, is described in Sec. 2.1. In Sec. 2.2, video decoding using the texture synthesis module is described. In Sec. 3, the experimental results are presented.

2 Hierarchical Texture Synthesizer

The hierarchical texture synthesis algorithm presented in this paper is a non-parametric approach inspired by the work by Kwatra et al. [1]. In this paper, Kwatra's algorithm is improved and adapted to a content-based video coding scenario. By selecting a non-parametric synthesis approach as baseline solution for our video codec, we opt for good video quality at the decoder and accept the potential drawback of increased side information compared to parametric algorithms.

2.1 Encoder

At the encoder, homogeneous texture segments are identified by the texture analyzer, which generates a mask sequence to delimit the identified texture clusters. Inhomogeneous texture areas are coded using H.264/MPEG4-AVC, while the others are synthesized at the decoder.

The masks are transmitted to the decoder by coding the synthesis macroblocks as skipped macroblocks [4]. That is, no transformation coefficients, residual error, nor motion vectors are transmitted for these macroblocks. However, a special skipped mode (TS mode) must be defined to distinguish between the H.264/MPEG4-AVC

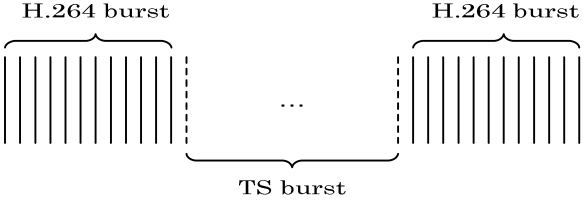


Fig. 1. GoP structure of the proposed video coding framework

skip and the texture synthesis skip. The TS mode is enabled with a few bits for each macroblock, which is reasonably compact side information. This is yet achieved by transferring higher computational complexity to the decoder.

The GoP structure of the coding scenario presented in this paper is depicted in Fig. 1. As can be seen, a burst of images coded using H.264/MPEG4-AVC is transmitted before and after the time interval that is to synthesize either partially or totally. The H.264/MPEG4-AVC coded frame bursts are used as texture examples for synthesis as is explained in the next section. Further eight motion parameters per burst frame are required at the decoder in case of a moving camera scenario as will be discussed in Sec. 2.2.

2.2 Decoder

Filling a Synthetic 3-D Texture

The texture synthesis approach presented in this paper is a patch-based algorithm. That is, the synthetic texture is generated patch-wise in opposition to sample-wise synthesis methods [5].

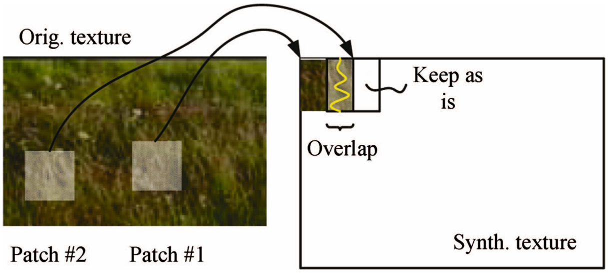


Fig. 2. 2-D representation of synthetic texture filling

A patch can be seen as a cuboid that represents a sub-volume of the total 3-D spatio-temporal volume to synthesize. A typical patch size can be given as 32x32x20, where a reduced patch size of 20 frames is used in temporal direction to minimize the length of the required H.264/MPEG4-AVC burst (cp. Fig. 1). The patches are inserted in an overlapping manner into the synthesis volume for reasons that will be explained in the next section. The overlap between the patches is typically chosen to be 16x16x10.

The first patch is selected at random from the example textures. The overlap volume (cp. Fig. 2) is then matched with the texture examples using an adequate measure. In this work, the mean squared error (MSE) measure is applied to the luminance channel of both the overlap volume and the example texture at full resolution. The continuation patch is selected at one of the locations of the example texture minimizing the MSE. That is, to avoid monotony or exact reproduction of the original texture, one of the best candidates for continuation is selected at random for effective synthesis. This process is pursued until the synthetic texture volume is completely filled.

Patch Adjustment using Graph Cuts

Once a continuation patch has been identified in the original texture, the probability of spurious edge occurrences at transitions between old and continuation (e.g. first and second) patch must be minimized. For that, an irregular shaped boundary is generated by using a graph cut approach [6] (cp. Fig. 2). The minimum cost path or cut from one end of the overlap volume to the other is determined. The cut thereby determines which patch contributes samples at different locations in the overlap region.

$$E = \frac{\left\| \mathcal{RGB}(p_{11}) - \mathcal{RGB}(p_{12}) \right\| + \left\| \mathcal{RGB}(p_{21}) - \mathcal{RGB}(p_{22}) \right\|}{\left\| \mathscr{g}_d(p_{11}) \right\| + \left\| \mathscr{g}_d(p_{21}) \right\| + \left\| \mathscr{g}_d(p_{12}) \right\| + \left\| \mathscr{g}_d(p_{22}) \right\|} \tag{1}$$

A cut between any sample p_1 of the overlap volume and any adjacent sample p_2 yields to the costs E introduced by Kwatra et al. [1] and formulated in (1). The numerator of (1) corresponds to the color difference between the sample pair p_1 and p_2.

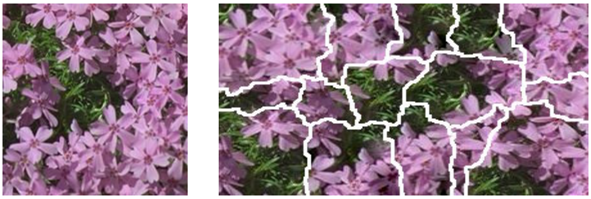


Fig. 3. Typical synthesis result given patch adjustment using graph cuts (right), example texture (left)

The difference is thereby measured in both old and continuation patch. p_{11} corresponds to sample p_1 in patch #1 (e.g. old patch), while p_{12} corresponds to sample p_1 in patch #2 (e.g. continuation patch). The same writing convention applies to sample p_2. $\mathcal{RGB}()$ corresponds to the RGB coordinates at the location indicated in the brackets. The denominator of (1) takes into account that false boundaries are less visible in high frequency than in low frequency regions. This is done by determining the gradients $\mathscr{g}_d()$ at locations p_1 and p_2 along direction d, where d can be x, y or t. $\mathscr{g}_d()$ is measured in the luminance channel of the overlap volumes.

The determination of the optimal cut, given the overlap volume and the corresponding costs, is based on the algorithm by Boykov et al. [6]. As a summary, it can be said that given E, the cut is typically chosen through high frequency regions, if any, to better dissimulate subjectively annoying discontinuities. A 2-D synthesis result using graph cuts for patch adjustment is depicted in Fig. 3 The cuts generated at patch transitions are shown as white irregular lines in the synthesis texture.

Hierarchical Texture Synthesis

Although graph cut produces the best possible seam given two overlapping patches, visible artefacts can still be generated when no good match exists for the given overlapping region of the old patch. This can for instance be the case if the old patch was chosen at the edge of the example texture. For that, a new hierarchical texture synthesis approach is proposed that allows for implicit smoothing of disturbing seams. The key feature of our proposal is that good cuts are not smoothed, while bad cuts are. Note that Kwatra et al. [1] use two different smoothing techniques, feathering and multiresolution splining, or do nothing in case the synthesis result is satisfactory. Which of the three options is used is decided using the trial and error method.

In the hierarchical texture synthesis approach, a Laplacian pyramid is built for each overlap region. The graph cut algorithm is first applied at the tip of the pyramid, that is, to the low frequency band of the signal. The synthetic texture, obtained after the cut, is filtered with a bandpass filter matching the frequency band of the current level of the pyramid. This is done to remove inappropriate frequencies that might have been introduced through synthesis.

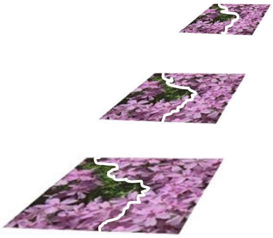


Fig. 4. Principle of hierarchical texture synthesis

This process is repeated at the downwards adjacent pyramid levels until the lowest level has been reached (cp.Fig. 4). Note that the cut at a given pyramid level is constrained by the cut at the previous level. That is, cut variations in the current plane are forced to lie within the 8-neighborhood of the interpolated previous cut. This is done to minimize the effect of noise on the cut procedure. Finally, the synthetic overlap region is obtained by fusing the cut frequency bands according to the reconstruction rules of the Laplacian pyramid. Note that the approach presented in this paper can be used for 2-D textures as well as for video textures with local motion (water, smoke, etc.) or for stiff video textures (brick wall, flowers, etc.).

Temporal Frame Alignment

The algorithm presented up to this point works only given a static camera. It can be extended to a more generic approach by temporally aligning the H.264/MPEG4-AVC bursts. This can be done w.r.t. the median of the total time interval concerned by the synthesis including the two H.264/MPEG4-AVC bursts. The original frames must be used at the encoder for this operation. For this purpose, an approach based on dense motion fields is implemented. Robust statistics are operated on the estimated motion vectors to derive the apparent camera motion and compensate it.

Let F_{tref} and $F_{tref\pm\alpha}$ be two frames of a video sequence. Then the dense motion field between the two is first estimated using the approach by Black and Anandan [7]. The samples belonging to background, i.e. regions without local motion activity, and thus underlying only global camera motion are determined using robust statistics, namely M-estimation [8]. The latter is an iterative model-fitting approach that detects outliers (non-background samples) within a dataset a posteriori and without any prior knowledge of outlier characteristics. The observations are a set of motion vectors in our specific framework, while outliers can be seen as motion vectors that reveal different motion properties than the inliers. Motion homogeneity is defined w.r.t. the perspective motion model [8]. That is, the observed motion field [7] is approximated using above-mentioned model. The perspective motion model (2) was selected due to its ability to describe translation, rotation, and scaling of a planar patch in 3-D as we assume this geometry for our synthesis textures. In (2), (x',y') represent the warped coordinates of the original sample (x,y), while $a_1,\ldots,a_8$ are the eight model parameters.

$$\begin{aligned} x' &= [(a_1 + a_3 x + a_4 y) / (1 + a_7 x + a_8 y)] \\ y' &= [(a_2 + a_5 x + a_6 y) / (1 + a_7 x + a_8 y)] \end{aligned} \tag{2}$$

The M-estimator minimizes the influence of outliers on the model optimization by penalizing motion vectors yielding high modeling costs. The cost function is thereby defined as the deviation between the observed [7] and the modelled (2) dense motion field as shown in (3).

$$E_M = |v_x - \omega_x| + |v_y - \omega_y| \tag{3}$$

In (3), (v_x,v_y) represent the observed, while (ω_x,ω_y) are the modelled motion vectors. After the global motion parameters $a_1,\ldots,a_8$ have been determined through M-estimation, $F_{tref\pm\alpha}$ is warped towards F_{tref} achieving temporal alignment in that way. The range of α ($\alpha \in \mathcal{N}$) depends on the type of camera operations in the video sequence: Large α ranges can be used for slow global motion, while smaller ranges must be used for fast motion.

After the alignment has been done, the texture synthesis is operated as described in the previous sections, which results in a synthetic texture w.r.t. t_{ref}. That is, each synthesized frame has to be warped towards the genuine time instant by using the inverse mapping of (2), which can be easily derived from the same equation. The temporal alignment, required in a moving camera scenario, yields the necessity to transmit additional side information, i.e. a motion parameter set $a_1,\ldots,a_8$ for each frame of the H.264/MPEG4-AVC burst.

3 Experimental Results

The experiments are conducted with three test sequences. Two of these correspond to a static camera scenario, while the third corresponds to a moving camera situation. A key frame of each sequence is depicted in Fig. 5. The two static camera sequences, ocean and grass, are synthesized using the hierarchical texture synthesis approach presented in this paper. The results are available under http://ip.hhi.de/imagecom_G1/HierarchicalTS.htm.

For the evaluation of the moving camera sequence, namely coast guard, the texture synthesizer is integrated into an H.264/MPEG4-AVC video codec. The following set-up is used for the codec. An unsynthesized burst of 21 frames is arranged at the beginning of the sequence, where the first frame of the burst is an I frame and the remaining frames are P frames. The 13 following GOPs which are composed of five B pictures and a P picture are partially synthesized (only water).

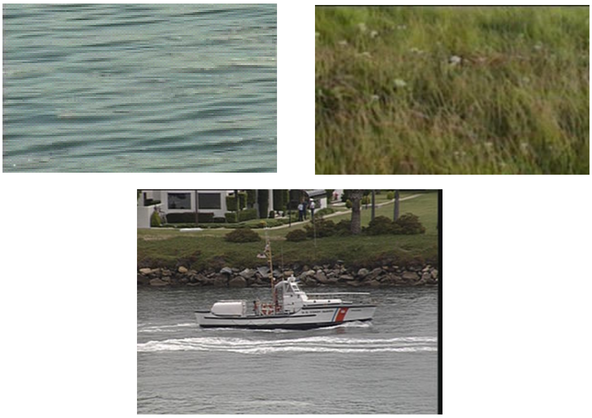


Fig. 5. Key frames of the test sequences ocean (top left), grass (top right), and coast guard (bottom)

One reference picture for each P picture, CABAC (entropy coding method), rate distortion optimization, 30 Hz progressive video at CIF resolution are also used. This setting results in a video sequence of a total length of 101 frames, as the video codec requires an additional P frame at the end of the sequence. Two configurations are used for the video codec without our approach. The first one is identical to the configuration described above, while the second one is entirely composed of GOPs of 5 B frames and a P frame.

It is found that better visual results can be achieved at the same bit rate (2661 kbps) for the H.264/MPEG4-AVC video codec with our approach compared to the codec without our algorithm. The decoding results can be down-loaded at the web-page mentioned above.

4 Conclusions and Future Work

In this paper, a hierarchical texture synthesis approach for content-based video coding is presented. It is shown that good video quality can be achieved at the decoder output for a large variety of video textures like water, grass etc. Compared to H.264/MPEG4-AVC, significant quality improvements can be achieved, at a constant bit rate, for video sequences containing such textures, as they are usually costly to code. The identification of continuation patches in the example texture will be improved to increase robustness of matches against noise. This can for instance be done by using a hierarchical matching procedure, instead of a full search at the highest signal resolution.

References

1. Kwatra, V., et al.: Graphcut Textures: Image and Video Synthesis using Graph Cuts. Proc. SIGGRAPH (2003) 277-286
2. Ndjiki-Nya, P., et al.: Improved H.264 Coding using Texture Analysis and Synthesis. Proc. ICIP (2003)

3. Dumitraş, A., and Haskell, B. G.: An Encoder-Decoder Texture Replacement Method with Application to Content-Based Movie Coding. IEEE Trans. on CSVT, Vol. 14, No. 6 (2004) 825-840
4. ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC: Advanced Video Coding for Generic Audiovisual Services. (2003)
5. De Bonet, J. S.: Multiresolution Sampling Procedure for Analysis and Synthesis of Texture Images. Proc. SIGGRAPH (1997) 361-368
6. Boykov, Y., et al.: Fast Approximate Energy Minimization via Graph Cuts. Proc. ICCV (1999) 377-384
7. Black, M. J., and Anandan, P.: The Robust Estimation of Multiple Motions: Parametric and Piecewise-smooth Flow Fields. Computer Vision and Image Understanding, Vol. 63, No. 1 (1996) 75-104
8. Ohm, J.-R.: Multimedia Communication Technology. Springer-Verlag, Berlin Heidelberg New York (2004)

An Intermediate Expressions' Generator System in the MPEG-4 Framework

Amaryllis Raouzaiou, Evaggelos Spyrou, Kostas Karpouzis, and Stefanos Kollias

Image, Video and Multimedia Systems Laboratory,
School of Electrical and Computer Engineering,
National Technical University of Athens,
Athens, Greece
{araouz, espyrou, kkarpou}@image.ntua.gr,
stefanos@cs.ntua.gr

Abstract. A lifelike human face can enhance interactive applications by providing straightforward feedback to and from the users and stimulating emotional responses from them. An expressive, realistic avatar should not "express himself" in the narrow confines of the six archetypal expressions. In this paper, we present a system which generates intermediate expression profiles (set of FAPs) combining profiles of the six archetypal expressions, by utilizing concepts included in the MPEG-4 standard.

1 Introduction

Research in facial expression analysis and synthesis has mainly concentrated on archetypal emotions. In particular, sadness, anger, joy, fear, disgust and surprise are categories of emotions that attracted most of the interest in human computer interaction environments. Very few studies [1] have appeared in the computer science literature, which explore non-archetypal emotions. This trend may be due to the great influence of the works of Ekman [4] and Friesen who proposed that the archetypal emotions correspond to distinct facial expressions which are supposed to be universally recognizable across cultures. On the contrary psychological researchers have extensively investigated a broader variety of emotions. An extensive survey on emotion analysis can be found in [9]. An expressive, realistic avatar should not "express himself" in the narrow confines of the six archetypal expressions. Intermediate expressions ought to be a part of synthesizable expressions in every possible application (online gaming, e-commerce, interactive TV etc).

Moreover, the MPEG-4 indicates an alternative way of modeling facial expressions and the underlying emotions, which is strongly influenced from neurophysiological and psychological studies (FAPs). The adoption of token-based animation in the MPEG-4 framework [6] benefits the definition of emotional states, since the extraction of simple, symbolic parameters is more appropriate to synthesize, as well as analyze facial expression and hand gestures.

In this paper we describe a system which has as output profiles of intermediate expressions, i.e. group of FAPs accompanied with FAP intensities - the actual ranges of variation, which if animated create the requested expression, taking into account

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 129 – 136, 2006.
© Springer-Verlag Berlin Heidelberg 2006

results of Whissel's study [9]. These results can then be applied to avatars, so as to convey the communicated messages more vividly than plain textual information or simply to make interaction more lifelike.

2 MPEG-4 and Emotion Representation

In the framework of MPEG-4 standard [8], parameters have been specified for Face and Body Animation (FBA) by defining specific Face and Body nodes in the scene graph. The goal of FBA definition is the animation of both realistic and cartoonist characters. Thus, MPEG-4 has defined a large set of parameters and the user can select subsets of these parameters according to the application, especially for the body, for which the animation is much more complex. The FBA part can be also combined with multimodal input (e.g. linguistic and paralinguistic speech analysis).

As far as facial animation is concerned, MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for FAPs definition. The FAP set contains two high-level parameters, visemes and expressions. In particular, the Facial Definition Parameter (FDP) and the Facial Animation Parameter (FAP) set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, eliminating the need for specifying the topology of the underlying geometry, through FDPs, and the animation of faces reproducing expressions, emotions and speech pronunciation, through FAPs. By monitoring facial gestures corresponding to FDP and/or FAP movements over time, it is possible to derive cues about user's expressions and emotions. Various results have been presented regarding classification of archetypal expressions of faces, mainly based on features or points mainly extracted from the mouth and eyes areas of the faces. These results indicate that facial expressions, possibly combined with gestures and speech, when the latter is available, provide cues that can be used to perceive a person's emotional state.

The obvious goal for emotion analysis applications is to assign category labels that identify emotional states. However, labels as such are very poor descriptions, especially since humans use a daunting number of labels to describe emotion.

Psychologists have examined a broader set of emotions [1], but very few of the studies provide results which can be exploited in computer graphics and machine vision fields. One of these studies, carried out by Whissel [9], suggests that emotions are points in a space spanning a relatively small number of dimensions, which seem to occupy two axes: *activation* and *evaluation* (*Figure 1*).

- *Valence* (Evaluation level): the clearest common element of emotional states is that the person is materially influenced by feelings that are "valenced", i.e. they are centrally concerned with positive or negative evaluations of people or things or events. The link between emotion and valencing is widely agreed (horizontal axis).
- *Activation* level: research has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e. the strength of the person's disposition to take some action rather than none (vertical axis).

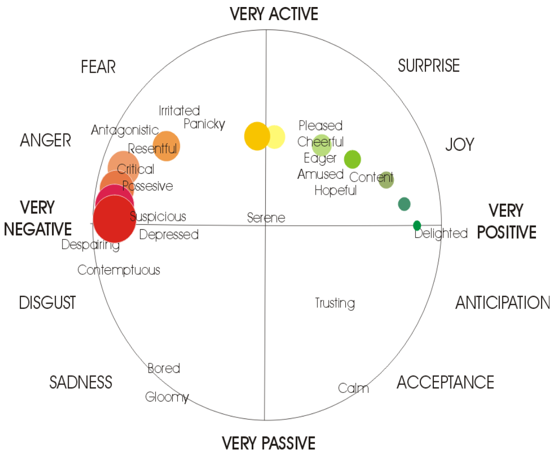


Fig. 1. The Activation – emotion space

A surprising amount of emotional discourse can be captured in terms of activation-emotion space. Perceived full-blown emotions are not evenly distributed in activation-emotion space; instead they tend to form a roughly circular pattern. In this framework, identifying the center as a natural origin has several implications. Emotional strength can be measured as the distance from the origin to a given point in activation-evaluation space. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation.

3 Intermediate Expressions

The limited number of studies, carried out by computer scientists and engineers [1], dealing with emotions other than the archetypal ones, lead us to search in other subject/discipline bibliographies. Psychologists examined a broader set of emotions [1], but very few of the corresponding studies provide exploitable results to computer graphics and machine vision fields, e.g. Whissel's study suggests that emotions are points in a space (*Figure 1*) spanning a relatively small number of dimensions, which in a first approximation, seem to occupy two axes: *activation* and *evaluation*, as shown in Table 1. *Activation* is the degree of arousal associated with the term, with terms like *surprised* (around 3) representing high activation, and *sad* (around -2) representing low *activation*. *Evaluation* is the degree of pleasantness associated with the term, with *angry* (at -3.0) representing the negative extreme and *delighted* (at 2.9) representing the positive extreme. From the practical point of view, *evaluation* seems to express internal feelings of the subject and its estimation through face formations is intractable. On the other hand, *activation* is related to facial muscles' movement and can be easily estimated based on facial characteristics.

Table 1. Selected Words from Whissel's Study

	Activation(a)	*Evaluation(e)*
Terrified	2.8	-2.0
Afraid	1.4	-2.0
Worried	0.4	-2.1
Angry	1.3	-3.0
Surprised	3	2.5
Sad	-2.0	-1.7
Depressed	-0.3	-2.5
Suspicious	0.2	-2.8
Delighted	0.7	2.9

The synthesis of intermediate expressions is based on the profiles of the six archetypal expressions [7].

As a general rule, one can define six general categories, each one characterized by an archetypal emotion. From the synthetic point of view, emotions that belong to the same category can be rendered by animating the same FAPs using different intensities. For example, the emotion group *fear* also contains *worry* and *terror* [7]; these two emotions can be synthesized by reducing or increasing the intensities of the employed FAPs, respectively. In the case of expression profiles, this affects the range of variation of the corresponding FAPs which is appropriately translated.

Creating profiles for emotions that do not clearly belong to a universal category is not straightforward. Apart from estimating the range of variations for FAPs, one should first define the FAPs which are involved in the particular emotion.

One is able to synthesize intermediate emotions by combining the FAPs employed for the representation of universal ones. In our approach, FAPs that are common in both emotions are retained during synthesis, while emotions used in only one emotion are averaged with the respective neutral position. In the case of mutually exclusive FAPs, averaging of intensities usually favors the most exaggerated of the emotions that are combined, whereas FAPs with contradicting intensities are cancelled out.

Below we describe the rules used by our system to merge profiles of archetypal emotions and create profiles of intermediate ones:

Let $P_{A_1}^{(k)}$ be the *k-th* profile of emotion A_1 and $P_{A_2}^{(l)}$ the *l-th* profile of emotion A_2. Let $X_{A_1,j}^{(k)}$ and $X_{A_2,j}^{(l)}$ be the ranges of variation of FAP F_j involved in $P_{A_1}^{(k)}$ and $P_{A_2}^{(l)}$ respectively. Additionally, $\omega = \tan^{-1}\left(\frac{a}{e}\right)$, ω_{A_1}, ω_{A_2} and ω_I, $\omega_{\mathrm{A}_1} < \omega_I < \omega_{\mathrm{A}_2}$, a_{A_1}, a_{A_2} and a_I are the values of the *activation* parameter and e_{A_1}, e_{A_2} and e_I the values of the *evaluation* parameter for emotion words A_1, A_2 and I respectively, obtained from Whissel's study [9]. The following rules are applied in order to create a profile $P_I^{(m)}$ for the intermediate emotion I:

Rule 1: $P_I^{(m)}$ includes FAPs that are involved either in $P_{A_1}^{(k)}$ or $P_{A_2}^{(l)}$.

Rule 2: If F_j is a FAP involved in both $P_{A_1}^{(k)}$ and $P_{A_2}^{(l)}$ with the same sign (direction of movement), then the range of variation $X_{I,j}^{(k)}$ is computed as a weighted translation of $X_{A_1,j}^{(k)}$ and $X_{A_2,j}^{(l)}$ in the following way: (i) the translated ranges of variations $t(X_{A_1,j}^{(k)}) = \frac{a_I}{a_{A_1}} X_{A_1,j}^{(k)}$ and $t(X_{A_2,j}^{(k)}) = \frac{a_I}{a_{A_2}} X_{A_2,j}^{(k)}$ of $X_{A_1,j}^{(k)}$ and $X_{A_2,j}^{(l)}$ are computed, (ii) the centers $c_{A_1,j}^{(k)}$ and $c_{A_2,j}^{(k)}$ of $t(X_{A_1,j}^{(k)})$ and $t(X_{A_2,j}^{(k)})$ are the same as those of $X_{A_1,j}^{(k)}$ and $X_{A_2,j}^{(l)}$, (iii) the lengths $s_{A_1,j}^{(k)}$ and $s_{A_2,j}^{(k)}$ of $t(X_{A_1,j}^{(k)})$ and $t(X_{A_2,j}^{(k)})$ are computed using the relation $s_{Ai,j}^{(k)} = \frac{1}{3} t(X_{A_1,j}^{(k)})$, (iv) the length of $X_{I,j}^{(k)}$ is

$$s_{I,j}^{(m)} = \frac{\omega_I - \omega_{A_1}}{\omega_{A_2} - \omega_{A_1}} s_{A_1,j}^{(k)} + \frac{\omega_{A_2} - \omega_I}{\omega_{A_2} - \omega_{A_1}} s_{A_2,j}^{(l)}$$ and its midpoint is

$$c_{I,j}^{(m)} = \frac{\omega_I - \omega_{A_1}}{\omega_{A_2} - \omega_{A_1}} c_{A_1,j}^{(k)} + \frac{\omega_{A_2} - \omega_I}{\omega_{A_2} - \omega_{A_1}} c_{A_2,j}^{(l)}$$

Rule 3: If the F_j is involved in both $P_{A_1}^{(k)}$ and $P_{A_2}^{(l)}$ but with contradictory sign (opposite direction of movement), then the range of variation $X_{I,j}^{(k)}$ is computed by $X_{I,j}^{(m)} = \frac{a_I}{a_{A_1}} X_{A_1,j}^{(k)} \cap \frac{a_I}{a_{A_2}} X_{A_2,j}^{(l)}$. In case where $X_{I,j}^{(k)}$ is eliminated (which is the most possible situation) then F_j is excluded from the profile.

Rule 4: If the F_j is involved only in one of $P_{A_1}^{(k)}$ and $P_{A_2}^{(l)}$ then the range of variation $X_{I,j}^{(k)}$ will be averaged with the corresponding of the neutral face position, i.e., $X_{I,j}^{(m)} = \frac{a_I}{2 * a_{A_1}} X_{A_1,j}^{(k)}$ or

$$X_{I,j}^{(m)} = \frac{a_I}{2 * a_{A_2}} X_{A_2,j}^{(l)}$$

4 Intermediate Expressions' Generator System

The proposed system (*Figure 2*) has as input the user request, i.e. the parameters *a* and *e* of Whissel's wheel of the desired intermediate expression or only the term of the intermediate expression (e.g. *depressed*) and the system uses the corresponding stored *a* and *e* values.

The main part of the system applies the above-mentioned rules and calculates the group of FAPs of every intermediate expression profile, accompanied by the corresponding values. This output has the exact form of the input of an MPEG-4 decoder, such as **GretaPlayer [5]**.

The profiles of each archetypal expression – every archetypal expression has more than one profile- are stored in the system and have the form of an array containing the ranges of variation for every one of the 68 FAPs of the MPEG-4 standard. However, only a subset of them can be extracted by the analysis procedure and thus can be used to create the profiles of archetypal expressions. The output of our system is also an array containing the essential information for an MPEG-4 decoder: a) a binary mask with ones representing the used FAPs and b) a value for every FAP lying near the center of the derived range of variation, picked randomly from the Gaussian distribution.

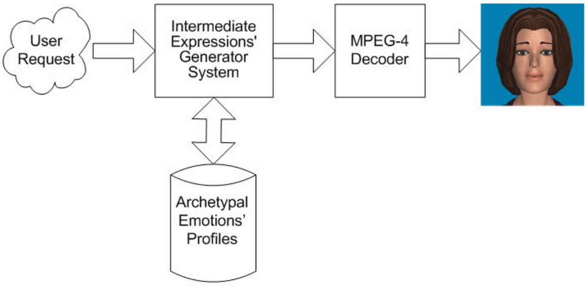

User Request
Intermediate Expressions' Generator System
MPEG-4 Decoder
Archetypal Emotions' Profiles

Fig. 2. Proposed System Architecture

5 Experimental Results

It should be noted that the profiles derived by the presented system, have to be animated for testing and correction purposes; the final profiles are those that are approved by experts, e.g. they present an acceptable visual similarity with the requested real emotion.

Table 2 presents the profiles of the basic emotions *fear* and *sadness* in the form they are stored in the system (see *Section IV*), omitting the columns that have zero values and the profile of the intermediate expression *depressed* in the same form.

Using the rules described above, *depression* (*Figures 3b, 3c*) and *guilt* (*Figure 5b*) is animated using *fear* (*Fig.3a, 5a*) and *sadness* (*Fig.3c, 5c*), *suspicious* (*Figure 4b*) using *anger* (*Fig. 4a*) and *disgust* (*Fig. 4c*). From *Figures 3b* and *3c, 3b* is approved and *3c* is rejected, after the judgment of an expert.

Table 2. Activation and evaluation measures used to create the profile for the emotion *depressed*

***Afraid* (1.4, -2.0):**

F_3	F_5	F_{19}	F_{20}	F_{21}	F_{22}	F_{31}	F_{32}	F_{33}	F_{34}	F_{35}	F_{36}
400	-240	-630	-630	-630	-630	260	260	160	160	60	60
560	-160	-570	-570	-570	-570	340	340	240	240	140	140

***Depressed* (-0.3, -2.5):**

F_3	F_5	F_{19}	F_{20}	F_{21}	F_{22}	F_{31}	F_{32}	F_{33}	F_{34}	F_{35}	F_{36}
160	-100	-110	-120	-110	-120	61	57	65	65	25	25
230	-65	-310	-315	-310	-315	167	160	100	100	60	60

***Sad* (-2.0, -1.7):**

F_3	F_5	F_{19}	F_{20}	F_{21}	F_{22}	F_{31}	F_{32}	F_{33}	F_{34}	F_{35}	F_{36}
0	0	-265	-270	-265	-270	30	26	0	0	0	0
0	0	-41	-52	-41	-52	140	134	0	0	0	0

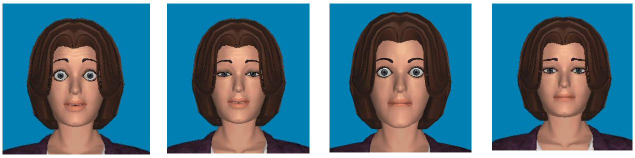

(a) (b) (c) (d)

Fig. 3. Profiles for (a) fear, (b-c) depressed (d) sadness

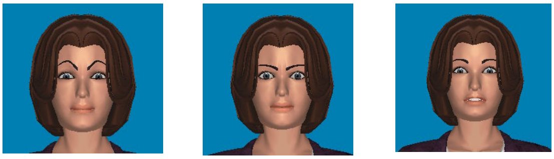

(a) (b) (c)

Fig. 4. Profiles for (a) anger, (b) suspicious (c) disgust

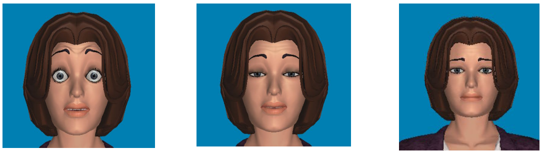

(a) (b) (c)

Fig. 5. Profiles for (a) fear, (b) guilt (c) sadness

6 Conclusions - Future Work

Expression synthesis is a great means of improving HCI applications, since it provides a powerful and universal means of expression and interaction. In this paper we presented a system which provides realistic intermediate facial expression profiles, utilizing concepts included in established standards, such as MPEG-4, which are widely supported in modern computers and standalone devices and making human-computer interaction more lifelike.

In the future, more profiles of intermediate emotions can be created by the combination of two expressions, not necessarily archetypal ones, while the same system with slight alterations may be used to generate gesture profiles of intermediate emotions.

References

1. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion Recognition in Human-Computer Interaction. IEEE Signal Processing Magazine (2001) 32-80
2. DeCarolis, B., Pelachaud, C., Poggi, I. and Steedman, M.: APML, A mark-up language for believable behavior generation, Life-Like Characters, Springer, 2004
3. EC TMR Project PHYSTA Report: Review of Existing Techniques for Human Emotion Understanding and Applications in Human-Computer Interaction (1998) http://www.image.ece.ntua.gr/physta/reports/emotionreview.htm
4. Ekman, P.: Facial expression and Emotion. Am. Psychologist, Vol. 48 (1993) 384-392
5. Hartmann, B., Mancini, M., Pelachaud, C.: Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis, *Computer Animation 2002*, pp. 111. 3, 6, 7
6. Preda, M. and Prêteux, F.: Advanced animation framework for virtual characters within the MPEG-4 standard, *Proc. of the Intl. Conference on Image Processing*. Rochester, NY, 2002.
7. Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S.: Parameterized facial expression synthesis based on MPEG-4. EURASIP Journal on Applied Signal Processing, Vol. 2002, No. 10. Hindawi Publishing Corporation (2002) 1021-1038
8. Tekalp, M., Ostermann, J.: Face and 2-D mesh animation in MPEG-4. Image Communication Journal, Vol.15, Nos. 4-5 (2000) 387-421
9. Whissel, C.M.: The dictionary of affect in language. In: Plutchnik, R., Kellerman, H. (eds): Emotion: Theory, research and experience: Vol 4, The measurement of emotions. Academic Press, New York (1989)

Coding with Temporal Layers or Multiple Descriptions for Lossy Video Transmission

Sila Ekmekci Flierl[1], Thomas Sikora[2], and Pascal Frossard[1]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL),
Signal Processing Institute & Institute for Telecommunications,
CH-1015 Lausanne, Switzerland
[2] Technical University Berlin, Institute for Telecommunications,
D-10587 Berlin, Germany

Abstract. In this paper, we compare temporal layered coding (TLC), as well as single-state coding (SSC), to multi-state video coding (MSVC) in the context of lossy video communications. MSVC is a Multiple Description Coding (MDC) Scheme where the video is coded into multiple independently decodable streams each with its own prediction process and state. The performance of these three coding schemes are analyzed at different loss rates and coding options, under the assumption that each packet contains the complete coded data for a frame, and the total bit rate is kept constant. To substitute the lost frames, MSVC employs state recovery based on motion compensated frame interpolation, whereas SSC and TLC repeat the last received frame. Results show that MSVC outperforms SSC and TLC for high motion sequences, and also for low motion sequences at high loss probabilities, due to increased state recovery ability of the system. Additionally, if one of the parallel channels of MSVC is in bad condition, unbalanced MSVC that allocates less bit rate to this channel, becomes favorable. Finally, increased error resilience with intra-GOB or frame update improves the system performance for high motion sequences at high loss rates, whereas for low motion sequences, intra updates are disadvantageous due to the penalty on the source coding quality.

1 Introduction

Video Communication over wireless networks and Internet is still a demanding issue due to long delays and packet losses which cause quality degradation. Multiple Description Coding [5] is a source coding technique used for transmission over error-prone channels. Two or more descriptions of the same source are generated which are mutually refining. If only one description is received the reconstruction distortion is D_1 or D_2. If both descriptions are received, however, a lower distortion D_0 is achieved. Multi-state video coding (MSVC) is a particular multiple description scheme where the video frames are split into two subsequences constituted of even and odd frames. Each subsequence can be encoded and decoded independently from each other. The advantage is twofold: 1- Even if one of the streams is lost the other one can still be decoded. 2- The lost frames can

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 137–145, 2006.
© Springer-Verlag Berlin Heidelberg 2006

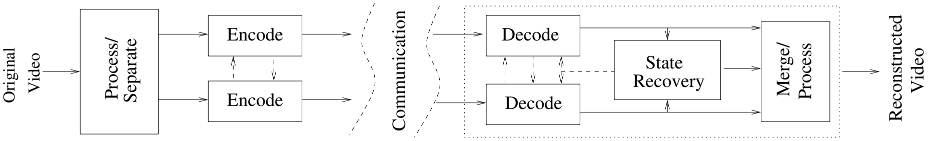


Fig. 1. Block Diagram of the MSVC System

be reconstructed by interpolation of their previous and next neighbors from the other subsequence (state recovery). Block diagram of the MSVC system is given in Figure 1. Reference [1] shows that if each frame is transmitted in a separate packet, MSVC outperforms SSC in recovering from single as well as burst errors. In this work, we compare the average performance of MSVC to SSC and also to Temporal Layered Coding (TLC) at the same total bitrate and at various channel loss rates (independent and uniformly distributed losses). Similar to MSVC, multiple bitstreams are generated in TLC [4]. Even if some portion of the bitstream is dropped due to channel problems, a reconstruction may still be possible with the received rest of the bitstream. However, in layered coding the reception of the base layer is mandatory for the decoding of the enhancement layer, contrarily to multiple description coding that enables independent decoding of the descriptions. Descriptions are in general mutually refining, while layers are hierarchically ordered, and thus natural candidates for differentiated protection.

In this paper, we consider streaming scenarios where multiple channels are available between the server and the client. MSVC uses both transmission paths, as well as TLC that separates the base and enhancement layers, where the enhancement layer contains each second frame coded as a B-frame. In SSC, the complete bitstream is sent over the same path. We investigate two cases: In the first case, the loss probabilities of the two paths are the same, i.e. $p_1 = p_2$, but the losses are independent from each other. In the second case, SSC is compared to MSVC, where one of the paths used for MSVC is lossless. We assume that each packet contains a frame and when a packet is lost, all information about the frame including the motion information is lost. In case of loss, SSC uses the last received frame to replace lost frames and MSVC implements state recovery based on motion compensated frame interpolation [1]. For all the comparisons, we target the same total bitrate R_T for the three coding methods. For MSVC, we investigate both balanced as well as unbalanced quantized MSVC ([3] and [2]). In the sequel, MSVC$_b$ denotes balanced quantized MSVC where the total bitrate R_T is allocated equally between the two streams considered, whereas MSVC$_u$ is the unbalanced MSVC where more bitrate is allocated to the more reliable channel. Additionally, we investigate the effect of GOB and frame intra updates on the three coding techniques. This way, we increase the number of resources for optimal rate allocation.

The paper is organized as follows: Section 2.1 presents the experimental setup, whereas several streaming scenarios are analyzed in section 2.2. Section 3 discusses the experimental results, and presents a series of heuristics particularly useful in the choice of an efficient coding strategy. Section 4 concludes the paper.

2 Comparative Analysis

2.1 Experimental Setup

H.264 codec is used for the experiments after modifying it to support the MSVC system. We consider two types of sequences: Foreman as a high motion sequence and Akiyo as a low motion sequence. The coding parameters (quantization step-sizes of intra and remaining frames, periods of GOB and frame updates and total bitrate R_T) are given in Tables 1, 2 and 3 for MSVC, SSC and TLC respectively, "A." denotes Akiyo and "F." stands for Foreman. The different rate allocations under consideration for MSVC$_u$ are given in Table 4. For all the comparisons, we target the same total bitrate R_T for all three coding methods (140 kbit/s for Foreman and 19 kbit/s for Akiyo). We considered the first 200 frames from each sequence.

The lossy transmission is simulated using random loss patterns. The average PSNR over all frames in each run is averaged over all loss patterns. 100 randomly generated loss patterns are used for each loss rate. MSVC uses Approach 2 from [1], that aims at maximizing the average frame PSNR by using interpolation from the past and future frames not only for lost frames, but also if the current frame PSNR can be increased through interpolation instead of using the received packet [3].

Table 1. MSVC+intra-updates, Coding Parameters

	QP I/P	**i.-GOB per.**	**i.-fr. per.**	**R_T [kbit/s]**
F.	17/17			158.21
F. i.-GOB	17/(20/21)	1		139.31
F. i.-fr.	17/23		9	140.82
A.	21/21			18.68

Table 2. SSC Coding Parameters

	QP I/P	**i.-GOB per.**	**i.-fr. per.**	**R_T [kbit/s]**
F.	16/16			137.28
F. i.-GOB	16/17	3		136.51
F. i.-fr.	16/17		30	133.88
A.	18/17			20.80

Table 3. TLC Coding Parameters

	QP I/P,B	**i.-GOB per.**	**i.-fr. per.**	**R_T [kbit/s]**
F.	14/14			147.95
F. i.-GOB	15/15	2		142.86
F. i.-fr.	15/15		15	143.60
A.	17/17			18.62

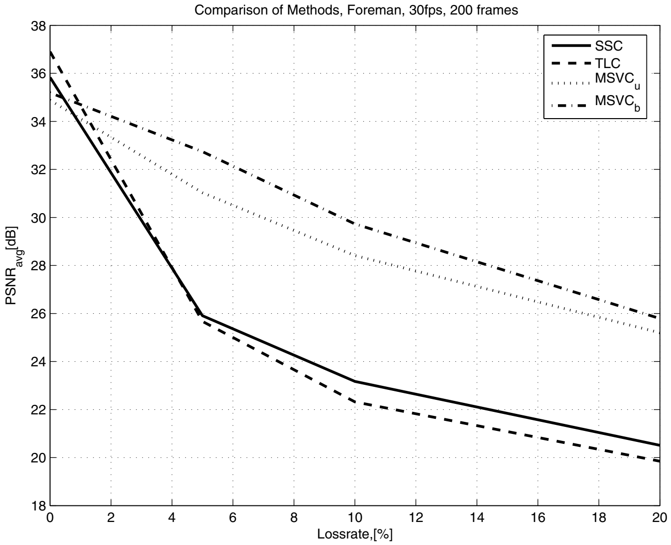
Table 4. Unbalanced rate allocation

	$\mathbf{R}_1$	$\mathbf{R}_2$
F.	111.88	27.48
F. i.-GOB	106.38	34.02
F. i.-fr.	83.90	55.68
A.	13.92	5.03

2.2 Results and Observations

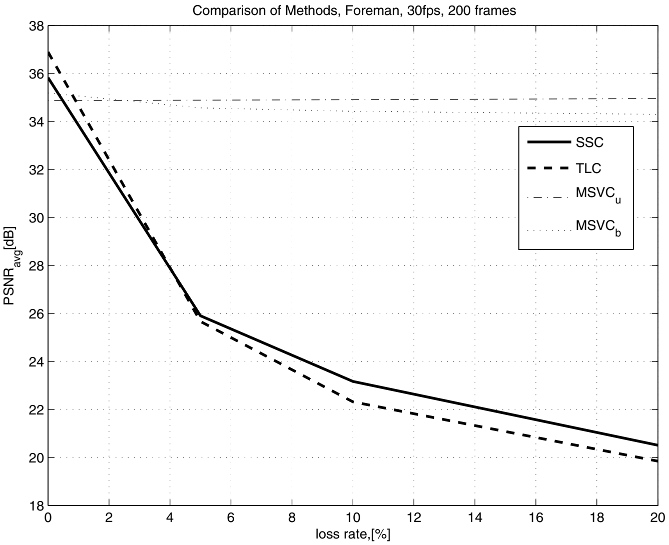
Figure 2 gives a comparison of the coding methods SSC, TLC and MSVC. MSVC outperforms SSC by 5 to 7 dB over the loss rate range when both of the channels have the same loss rate. This is a huge gain although we assumed that both channels are error prone. Moreover SSC outperforms TLC as the loss rate increases: at 20% loss rate, the gap between the two methods is about 0.8 dB. Figure 3 shows the case when the first channel used for MSVC is lossless whereas the second one has the same loss rate as the channel used by SSC. The probability that we catch a second channel with a better transmission condition is the main idea behind path diversity. At 20% loss rate, MSVC$_u$ outperforms SSC by 14 dB when $p_1 = 0\%$. The PSNR gap between MSVC$_u$ and MSVC$_b$ is about 1 dB at 20% loss rate, i.e.: unbalanced channels call for unbalanced rate allocations.

Figures 4 and 5 show the same comparisons for the low motion sequence Akiyo. Error concealment is easier due to low motion. Therefore SSC with repetition of the last received frame as concealment technique gives good performance in lossy environment and outperforms MSVC when loss rate is smaller than about 15%



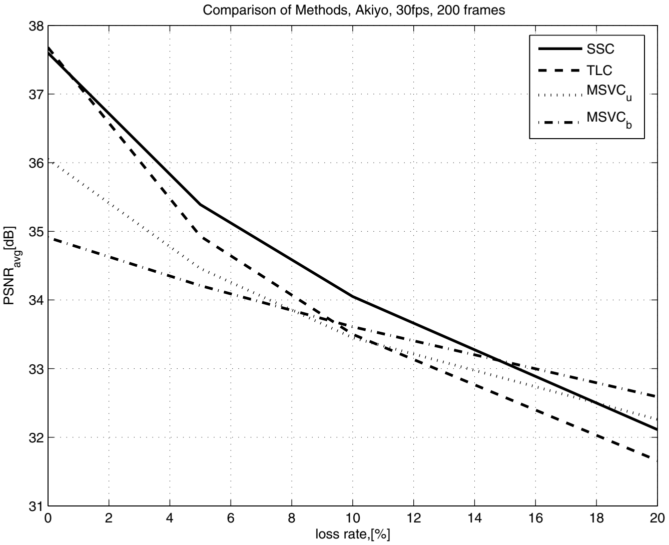
,

Fig. 2. Comparison of SSC, TLC, MSVC$_u$ and MSVC$_b$, all channels have the same loss rate, Foreman



,

Fig. 3. Comparison of SSC, TLC, MSVC$_u$ and MSVC$_b$, one of the MSVC channels is lossless, Foreman



,

Fig. 4. Comparison of SSC, TLC, MSVC$_u$ and MSVC$_b$, all channels have the same loss rate, Akiyo

as shown in Figure 4. But when loss rate increases beyond this limit, it is better to employ MSVC. Although unbalanced rate allocations are better at smaller loss rates (MSVC$_u$), larger loss rates require balanced rate allocations (MSVC$_b$). When the first channel is lossless, MSVC$_u$ performs always better than MSVC$_b$. Moreover, MSVC$_u$ performs 4dB better than SSC at 20% loss rate.

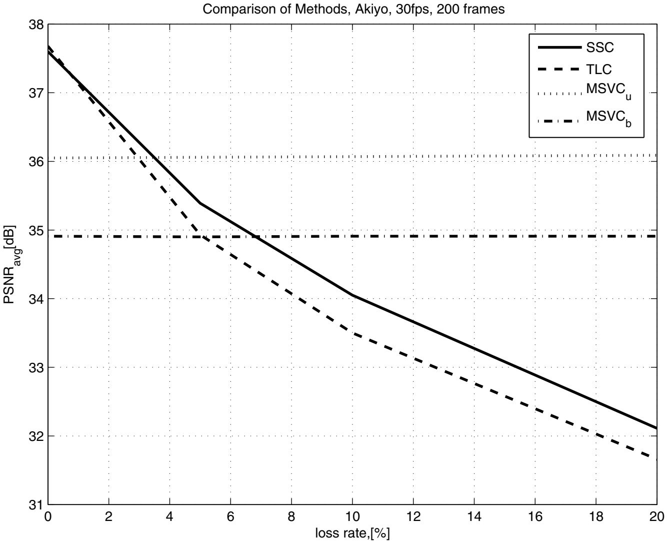


Fig. 5. Comparison of SSC, TLC, MSVC$_u$ and MSVC$_b$, one of the MSVC channels is lossless, Akiyo

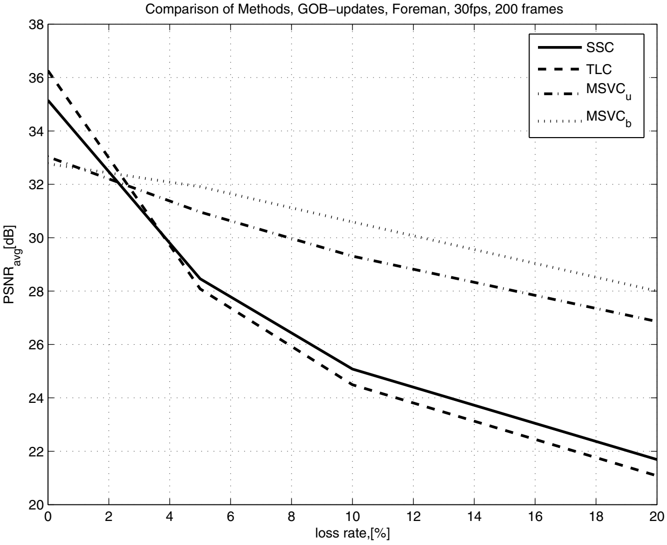
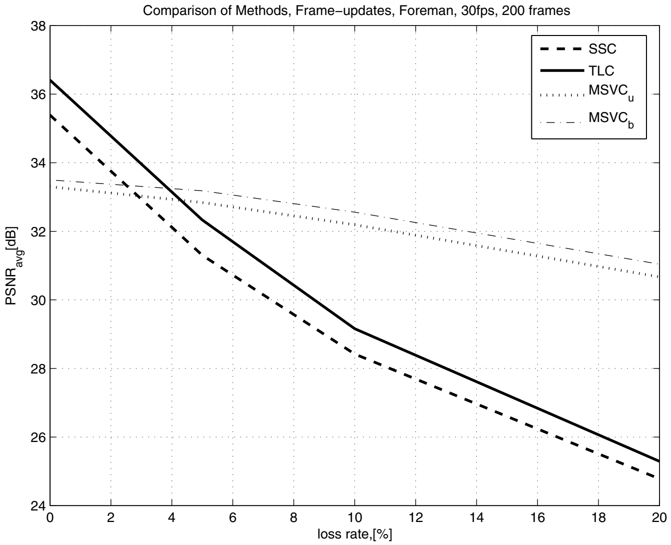


Fig. 6. Comparison of SSC, TLC, MSVC$_u$ and MSVC$_b$, all channels have the same loss rate, Foreman with GOB-intra-updates

In the next step, we compare the methods when intra-updates are used. Figures 6 and 7 show the cases with intra GOB- and frame updates for Foreman respectively. The threshold loss probability increases with the introduction of updates. All coding techniques profit from updates at high loss probabilities. The performance increase in SSC and TLC is larger than in MSVC. Using intra



,

Fig. 7. Comparison of SSC, TLC, MSVC$_u$ and MSVC$_b$, all channels have the same loss rate, Foreman with frame-intra-updates

updates for Akiyo is not a good idea, since the gain of motion compensation is very high and a small rate is available for coding (The corresponding figures are omitted here due to limited space). TLC performs best, since enhancement layer uses no updates. The differences between different methods are smaller for Akiyo. MSVC outperforms SSC at about 15% loss rate. Moreover, MSVC outperforms SSC when the first channel is lossless.

3 Discussion

For Foreman, when both of the channels have the same loss rate, MSVC outperforms SSC and TLC. The difference increases with increasing loss rate. But at lossless transmission there is a penalty for MSVC due to sequence splitting, i.e. increased temporal distance which decreases the prediction gain. Moreover introducing intra-updates increases the performance of MSVC as well as of other coding methods for high loss rate. But for lossless transmission, the performance drops due to the wasted bitrate for intra coding. For the same total bitrate R_T, intra updates give better performance than GOB intra updates. Additionally, we see that balanced loss probabilities call for balanced rate allocations.

For Akiyo, however, repetition of the last received frame in case of losses gives good results due to low motion. MSVC outperforms SSC only at high loss probabilities. Frame splitting in MSVC is disadvantageous due to the high cost of intra frames (the first frame of each subsequence is coded intra).

The slopes of distortion-loss rate curves for MSVC$_u$ and MSVC$_b$ are very small when one of the channels is lossless, as shown in Figures 3 and 5. The reason is that the average PSNR for the lossless received stream does not change with the loss rate, and that the lost frames in the lossy stream are reconstructed through interpolations from the lossless stream. Moreover, the slope for MSVC$_u$ is smaller than that for MSVC$_b$. Since more bitrate is allocated to the reliable channel (smaller quantization distortion), interpolation errors in case of losses are smaller. For Akiyo, the slopes of MSVC$_u$ and MSVC$_b$ are nearly zero, since frame interpolation gives always good results due to low motion.

The experiments show that if one of the channels is lossless, MSVC outperforms both SSC and TLC for both sequences. Even if both of the channels have the same loss probability, at high loss probabilities MSVC gives the best performance. The threshold loss probability for MSVC is dependent on the motion content of the sequence, e.g. about 0.5% for Foreman and 15% for Akiyo. The performance gap between SSC and MSVC is larger for high motion sequences.

4 Conclusions

To decide which coding technique and which coding options are to choose, the following factors are important: motion content of the sequence and loss rate of the channels (obtained by methods like channel probing etc.). Generally, MSVC is to be preferred at high loss probabilities. The threshold loss probability where MSVC outperforms SSC is higher for low motion sequences. Moreover, introduction of intra-updates increases the threshold loss probability, i.e. SSC and TLC profits more from intra updates than MSVC. For high motion sequences, MSVC combined with frame-intra-frames gives the best results at high loss probabilities. For low motion sequences, however, intra-updates decrease the system performance. Frame-intra-updates are more efficient than GOB-intra-updates in recovering from state errors.

In this paper, we compared MSVC to SSC and TLC at different loss rates and coding options. In each case, we targeted a constant total bitrate R_T to allow a fair comparison. Both for MSVC and TLC, we assumed that two independent channels are in use with independent loss patterns. We investigated both balanced and unbalanced operation for MSVC. In balanced operation half of the total bitrate is allocated to each stream, whereas in unbalanced case more bitrate is assigned to the first channel which is more reliable than the second one.

Further work will focus on joint optimization of redundancy, frame rate and also the quantization stepsize of the MSVC streams depending on the channel loss probabilities.

References

1. Apostolopoulos, J.: Reliable video communication over lossy packet networks using multiple state encoding and path diversity. Visual Communications and Image Processing (2001)
2. Ekmekci, S., Sikora, T.: Unbalanced quantized multi-state video coding:Potentials. Picture Coding Symposium (2004)

3. Ekmekci, S., Sikora, T.: Unbalanced quantized multiple description video transmission using path diversity. Electronic Imaging (2003)
4. Ghanbari, M.: Two-layer coding of video signals for vbr networks. IEEE Journal on Selected Areas in Communications **7** (1989).
5. Goyal, V. K.: Compression meets the network. IEEE Signal Processing Mag. **18** no. 5, (2001) 74–98

Fast Mode Decision Based on Activity Segmentation in H.264/AVC Encoding

Marcos Nieto, Luis Salgado, and Narciso García

Grupo de Tratamiento de Imágenes – E.T.S. Ingenieros de Telecomunicación,
Universidad Politécnica de Madrid, Spain
{mnd, lsa, narciso}@gti.ssr.upm.es

Abstract. The recent H.264 video coding standard has been developed to increase video coding efficiency, achieving better Rate-Distortion results than any other standard. This is mainly due to the new features included at the motion estimation process, as tree structured motion compensation and multiple reference frames. However, the motion estimation stage is the most expensive part of a video encoder, and these improvements greatly intensify this computational load. To solve this situation, in this paper we propose a fast mode decision scheme, that analyses the activity between frames and classify the picture into active and non-active areas. This classification is performed with and adaptive thresholding technique that minimizes the area set as active and the activity energy of the area set as non-active. For each macroblock, a sub-set of inter-prediction alternatives is selected according to the amount of active and non-active area inside it. Results show motion estimation time reduction of 38% to 51% while achieving Rate-Distortion values very similar to the ones obtained with the complete analysis of inter-prediction alternatives.

1 Introduction

The new international video coding standard H.264/AVC [1][2] has been developed to improve video coding efficiency, achieving half bit-rate and maintaining the same objective quality, compared to earlier standards such as MPEG-2 or recent MPEG-4. Motion Estimation (ME) is considered one of the most important parts of a video encoder to obtain a good Rate-Distortion (RD) performance [3]. Therefore, H.264/AVC has been designed to intensify this process by including some efficient new features like: multiple reference frames, tree structured motion compensation (MB subdivision into smaller blocks) and quarter-pixel motion vector accuracy.

For each macroblock (MB), a complete set of inter prediction alternatives is computed based on sub-dividing the MB into smaller blocks, down to sizes of 4x4 pixels, as it is shown in Fig. 1.

The selection of the best mode is usually taken performing RD optimization algorithms [3] applying the Lagrangian multiplier technique [4], by minimizing the function

$$J_m = D_{rec} + \lambda_m \cdot R_m \tag{1}$$

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 146 – 152, 2006.
© Springer-Verlag Berlin Heidelberg 2006

where D_{rec} is the sum of absolute difference (SAD) between the current MB and its reconstructed MB after quantization, λ_m is the Lagrangian multiplier and R_m are the bits to encode the MB header, the motion vector difference (MVD) and the residual DCT coefficients. This criterion provides the total RD coding cost for a MB, so the encoder just selects the mode that offers the minimum value of the function $J_{m.}$

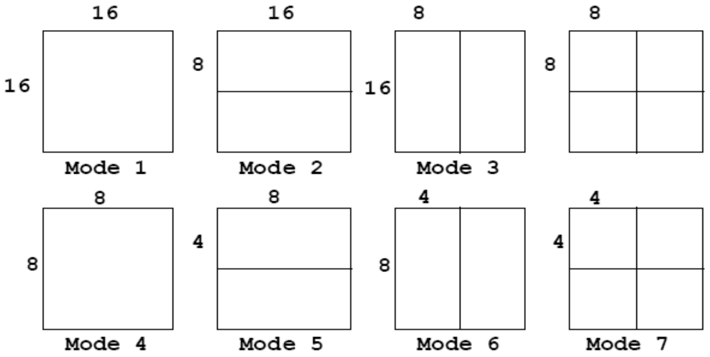


Fig. 1. Sub-partition block scheme of H.264/AVC

However, the complete analysis of inter prediction alternatives produces an enormous computational load that significantly increases the encoding time, avoiding H.264/AVC to be used in applications that requires fast and simple encoders.

To reduce this computational load, fast mode decision algorithms [5][6] have been proposed to reduce the number of inter-prediction alternatives to be computed for each MB, discarding those ones that seems to be redundant.

We introduce a new Fast Mode Decision algorithm that selects a sub-set of inter prediction alternatives, based on performing more coding effort, by selecting a large sub-set of modes, on those areas of the pictures that contain more activity energy and less coding effort on non-active areas, by selecting a small sub-set of modes.

The activity image of each frame is obtained as the difference picture between the current original frame, F_n, and the immediately previous reconstructed frame, F'_{n-1}. Within those activity images, areas are classified into active and non-active by applying an automatic thresholding technique that delivers as output an activity segmentation mask. The objective of this thresholding technique is to minimize the active area of the resulting segmentation mask (to achieve great computational load reduction) and the activity energy belonging to the areas of the activity image set as non-active (to preserve the RD cost as similar as possible to the one obtained with the analysis of the complete set of modes).

After the activity segmentation is performed, a sub-set of inter-prediction alternatives is selected for each MB to be computed at the ME process, according to the amount of active and no-active areas inside it. For a MB that mainly contains non-active areas, a small sub-set of modes is selected and therefore less coding effort is done. On the other hand, for MBs with active areas inside it, a larger sub-set of modes is selected and more coding effort is done. This way, the more non-active areas contained at the segmentation mask, the more MBs with non-active areas inside it, and therefore, the more ME time reduction achieved.

The main objective is to reduce the computational load of the ME process, so the fast mode decision scheme must not introduce its own computational overload that reduces the achieved time reduction at the ME process. Therefore, the algorithms that will be described in next paragraphs have been selected due to their simplicity and their unnoticeable computational overload for the ME process.

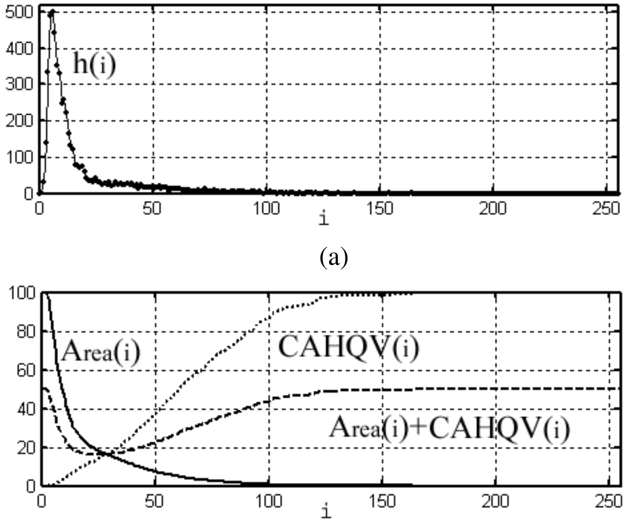
The paper is organized as follows: Section 2 shows the characteristics of the activity image; Section 3 describes the selection of the threshold that separates areas into active and non-active; Section 4 presents the algorithm that performs the selection of the sub-set of modes according to the amount of activity contained at each MB; and Section 5 and 6 show the results and conclusions achieved.

2 Activity Image

The resolution of the activity image is 4x4 times lower than the resolution of the sequence frames. Each pixel of the activity image is computed as the MAE between the corresponding block of the original frame, F_n, and the immediately previous reconstructed frame, F'_{n-1}.

$$MAE = \frac{1}{16}\sum_{i=0}^{3}\sum_{j=0}^{3}\left|B_{i,j}^{N} - B_{i,j}^{N-1}\right| \tag{2}$$

where B^N and B^{N-1} represent each 4x4 block belonging to F_n and F'_{n-1} respectively.



(a)

(b)

Fig. 2. (a) Unimodal histogram; (b) Area(i) and CAHQV(i), with the scaled sun function Area(i)+CAHQV(i)

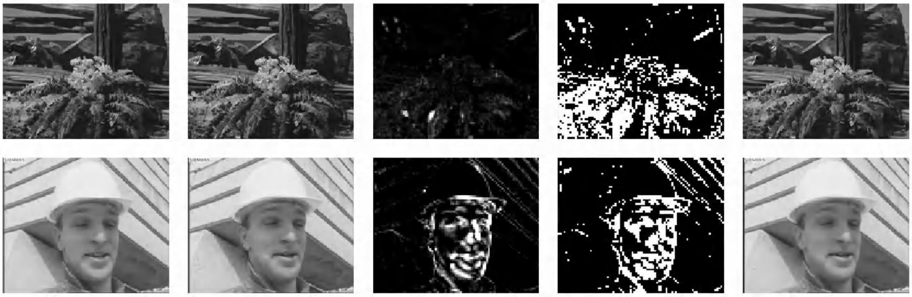


Fig. 3. Sequence *Tempete* (upper images), and *Foreman* (lower images): a) F'$_{n-1}$; b) F$_n$; c) Activity image with scaled values for visualization; d) Activity mask after thresholding, white areas are active areas and black areas are non-active areas; e) Reconstructed frame after mode decision, F'$_n$

This size is selected as 4x4 is the minimum sub-division block size permitted by H.264/AVC [1]. As a result, each 16x16 MB of F$_n$ contains 16 activity values corresponding to the MAE value of each 4x4 block that the MB contains. Therefore, the mode selection algorithm, described in Section 4, is simplified for each MB, as the amount of active or non-active areas inside each MB is reduced to 16 values.

The activity image is usually composed by specific areas where there are high activity values, whereas the main part of the image remains non-active. If there is activity distributed along the whole picture, this activity has usually lower values. As a consequence, typical histograms of these activity images show unimodal distributions, as it is shown in Fig. 2 (a). The main peak of the histogram is typically very close to the cero value, and the values corresponding to the tail of the histogram are those that contain the most relevant activity of the picture. Fig. 3 shows examples of the activity image generated for the two different sequences, *Tempete* and *Foreman*. In the case of the sequence *Foreman*, the activity is more important at some areas of the face while less activity is detected at the background.

Although there are significantly more number of 4x4 blocks with low activity values than with high activity values, their energy is not as large as the energy of the blocks with high activity. Therefore, the activity energy (or energy of the activity image) is clearly not concentrated around the main peak of the histogram but distributed along the tail of the histogram.

To study the energy distribution of the activity image the *cumulative activity histogram of quadratic values*, CAHQV(i), is built, defined as:

$$CAHQV(i) = 100\left[\left(\sum_{k=0}^{i} k^2 \cdot P(k)\right) \Big/ \left(\sum_{k=0}^{I} k^2 \cdot P(k)\right)\right] \tag{3}$$

where P(i) are the histogram probability values computed from the histogram values h(i):

$$P(i) = h(i) \Big/ \sum h(k) \tag{4}$$

and I is typically 255 for one byte level images. One example of *CAHQV(i)* is shown in Fig. 2 (b), where the curve shows an energy distribution with a soft slope, which indicates that the activity values that do not belong to the main histogram peak contain the most significant part of the energy of the picture.

3 Thresholding

To classify the activity image into active and non-active areas, a threshold is dynamically selected minimizing the sum of two parameters:

1) The amount of activity energy that the threshold sets as non-active. The objective here is to obtain the less possible RD deviation compared with the analysis of the complete set of inter prediction modes. The more energy set as non-active, the more RD deviation, as for non-active areas less coding effort is done selecting a smaller sub-set of inter-prediction alternatives.
2) The active area of the segmentation mask. Reducing the active area implies larger computational cost reduction, as for non-active areas less number of inter prediction alternatives are analyzed. This area, *Area(i)* is obtained as follows:

$$Area(i) = \left(\sum_{j=0}^{I} h(j) - \sum_{j=0}^{i} h(j) \right) \Bigg/ \sum_{j=0}^{I} h(j) \qquad (5)$$

This way, the threshold is obtained at the minimum value of the sum of the curves *CAHQV(i)* and *Area(i)* as it is shown in Fig. 2 (b), where the sum *Area(i)+CAHQV(i)* offers a minimum value near the intersection point of the curves (the sum function has been scaled for a better visualization).

After thresholding, the segmentation mask obtained is a binary image, as those shown in Fig. 3 (d) for *Tempete* (lower image) and *Foreman* (upper image): black regions represent non-active blocks and the white ones represent active blocks.

4 Mode Decision

Once the segmentation of the activity images has been performed, the active and non-active 4x4 blocks inside each MB are known, so the algorithm selects an appropriate sub-set of inter prediction alternatives in order to perform more coding effort on active areas and less coding effort on non-active areas.

The reason is that for those areas marked as non-active, the RD cost achieved through the selection of the best mode from the complete set of inter prediction modes is usually very similar to the RD cost achieved by any of the other modes. Therefore, if the number of modes to be computed at the ME process is appropriately reduced, the RD cost will have low deviation from the complete analysis of modes, while the computational load of the ME process will be dramatically reduced.

A block diagram of the sub-set selection algorithm is shown in Fig. 4. The flow chart describes the steps where modes are enabled (added to the sub-set of modes to be computed at the ME process) or left as disabled.

- If the MB contains 16 4x4 blocks marked as non-active, SKIP-Mode is selected and no more modes are computed for ME. Otherwise the algorithm continues:
- If there are some active blocks inside the MB, Modes 1, 2 and 3 are enabled.

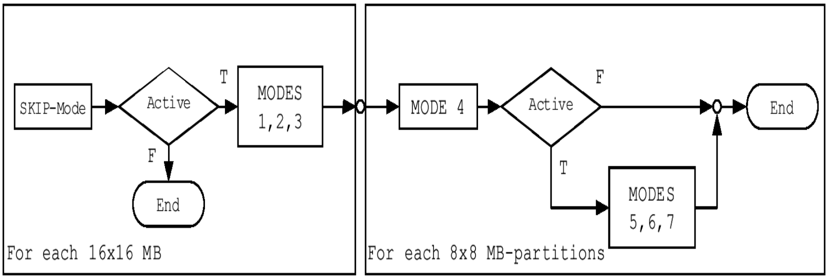


Fig. 4. Mode Decision Block Diagram for each MB, where modes are enabled: (SKIP-Mode), (MODES 1, 2, 3), (MODE 4) and (MODES 5, 6, 7); *Active* tests if there is any active 4x4 block inside the MB or the MB-partition under test

Additionally, for each 8x8 MB-partition:

- If its 4 4x4 blocks are marked as non-active, only Mode 4 (8x8) is selected for that MB-partition.
- If there are some active blocks inside it, Modes 5, 6 and 7 are enabled.

5 Results

To make experimental tests, the proposed fast mode decision algorithm has been implemented into the H.264/AVC reference software JM 9.3 [7].

Results with different standard test sequences are presented in Table 1, where ΔPSNR represents the difference of the obtained PSNR values between the encoder with the proposed fast mode decision algorithm and the reference encoder, ΔRate is analogously the difference of the output bit-rate, and the average reduction of computational load, ΔT, is computed as:

$$\Delta T(\%) = \left(\left(T_{sub-set} - T_{ref}\right) / T_{ref}\right) \cdot 100 \qquad (6)$$

Table 1. Rate-Distortion and Computational load reduction results

Sequence	ΔPSNR (dB)	ΔRate (%)	ΔT(%)
Coastguard (CIF)	-0.19	+2.72	-45.09
Container (CIF)	0.00	-0.07	-40.20
News (QCIF)	-0.04	-0.08	-51.31
Foreman (CIF)	-0.21	+2.38	-48.00
Suzie (QCIF)	-0.07	+0.55	-38.89
Claire (QCIF)	0.00	+0.05	-50.34
Carphone (QCIF)	-0.07	-0.52	-40.98
Silent (QCIF)	-0.04	+0.01	-45.07
Tempete (QCIF)	-0.37	+2.11	-43.85
Average	-0.11	+0.79	-44.86

where $T_{sub\text{-}set}$ is the achieved ME time with the proposed method, and T_{ref} is the ME time of the reference encoder (which computes all the possible modes for each MB). In these figures, the computation overload introduced by the algorithms presented in this paper is minimum, as they never involve more than 0.01% of the ME time.

As it is shown, the average ΔT achieved is –44.86 %, with the lowest value of 38,89% for *Suzie*, where the complex and fast movements imply high activity values distributed over very significant areas of the image.

The proposed method obtains excellent results, reducing the computational load and obtaining RD values very similar to the RD achieved with the analysis of the complete set of inter prediction alternatives. Average values shows PSNR deviation up to –0.11 dB and bit-rate deviation of +0.79% while subjective loss of quality is unnoticeable.

6 Conclusions

We have introduced a new fast mode decision scheme that allows great computational load reduction at the ME process with small RD deviation compared with the complete analysis of modes. This scheme contains an automatic thresholding technique that classifies the areas of each frame into *active* or *non-active*, selecting for active areas larger sub-set of modes, and small sub-sets for inactive areas.

Our results show excellent ME time reduction for a large variety of sequences, with small RD penalty while maintaining the same quality.

References

1. Wiegand, T., Sullivan, G. J.: Overview of the H.264/AVC video coding standard. IEEE Transactions on Circuits and Systems for Video Coding **13**, n. 7 (2003) 560-576.
2. Richardson, I. E. G.: H.264 and MPEG-4 Video Compression. Video Coding for Next-generation Multimedia. Ed. Wiley (2003).
3. Sullivan, G. J., and Wiegand, T.: Rate-distortion Optimization for Video Compression IEEE Signal Processing Mag. **15**, no. 6 (1998) 74-90.
4. Wiegand, T., and Girod, B.: Lagrange Multiplier Selection in Hybrid Video Coder Control. Proc. Int. Conf. on Image Processing, ICIP'01, Thessaloniki, Greece **3** (2001) 542-545,
5. Lee, J., Jeon, B.: Fast Mode Decision for H.264 with Variable Motion Block Sizes. ISCIS 2003, *Lecture Notes in Computer Science* **2869** (2003) 723-730.
6. Choi, I., Lee, J., and Jeon, B.: Efficient Coding Mode Decision in MPEG-4 Part-10 AVC/H.264 Main Profile. Proc. Int. Conf. on Image Processing, ICIP'04 (2004) 1141-1144.
7. H264/AVC Reference Software Model (JM 9.3): http://bs.hhi.de/~suchring/tml/index.html

Atom Position Coding in a Matching Pursuit Based Video Coder

Pierre Garrigues and Avideh Zakhor*

Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley, CA 94720
{garrigue, avz}@eecs.berkeley.edu

Abstract. In this paper, we propose a new scheme for position coding of the atoms within the Matching Pursuit algorithm as applied to the displaced frame difference (DFD) in a hybrid video encoder. We exploit the spatial and temporal coherence of the positions of the atoms in the DFD: the atoms tend to cluster in regions where the motion prediction fails. The location of these regions is correlated over time. We use a block-based technique in which the number of atoms inside each block is coded differentially with respect to the same block in the previous frame. The atom positions inside each block are coded using the previously proposed NumberSplit algorithm. We demonstrate the effectiveness of our proposed approach on a number of video sequences.

1 Introduction

The hybrid motion-compensated Discrete Cosine Transform (DCT) structure or a DCT-like integer transform structure are a part of nearly all existing video coding standards such as H.263, H.264/AVC and MPEG-1,2,4, and commercially available video compression systems. In each of these systems, a block-based motion model is used to predict each frame from a previously reconstructed frame, and the residual energy in each motion block is coded using the DCT. One of the problems with DCT is the blocking artifacts typically associated with all block-based transforms. To circumvent blocking artifacts, a number of non-block-based schemes based on expansion of signals on overcomplete set of basis functions have been proposed. One algorithm for such an expansion is Matching Pursuit (MP), originally proposed in the statistical literature [3], and later re-introduced to the signal processing community by Mallat and Zhang [6]. Application of MP to residual video coding was originally proposed in [9], and further developed in [1][2][4][8][5][7]. MP-based codecs have been found to result in significant PSNR and visual quality improvement over DCT-based schemes as they contain no blocking artifacts.

In analyzing the coding statistics of MP-based video codec in [1], we have empirically determined that a significant portion of the bits are devoted to atom position coding. Other sources of bit consumption include motion vectors, atom

* This work was supported by AFOSR contract F49620-00-1-0327.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 153–160, 2006.
© Springer-Verlag Berlin Heidelberg 2006

modulus coding and indices of the elements in the dictionary. Specifically, at higher bit rates, position coding represents up to 40% of the bit rate. Since the performance gap between MP and block-based techniques decreases at higher bit rates, it is conceivable to reduce this gap by more efficient position coding techniques.

In this paper, we propose a new position coding technique based on the exploitation of the spatial and temporal coherence of the atom positions. The outline of the paper is as follows. In Sections 2 and 3, we review the MP theory and existing position coding techniques respectively. Our new position technique and its application are discussed in Section 4. Finally, experimental results are included in Section 5.

2 Matching Pursuits Theory

The MP algorithm, as proposed by Mallat and Zhang [6], expands a signal onto an overcomplete dictionary of functions. For simplicity, the procedure can be illustrated for a 1-D time signal. Suppose we want to represent a signal $f(t)$ using basis functions from a dictionary set Γ. Individual dictionary functions are denoted as:

$$g_\gamma(t) \in \Gamma \tag{1}$$

where γ is an indexing parameter associated with a particular dictionary element. The decomposition begins by choosing γ to maximize the absolute value of the following inner product:

$$p = < f(t), g_\gamma(t) > . \tag{2}$$

We then say that p is an expansion coefficient for the signal onto the dictionary function $g_\gamma(t)$. A residual signal is computed as:

$$R(t) = f(t) - p g_\gamma(t). \tag{3}$$

This residual signal is then expanded in the same way as the original signal. The procedure continues iteratively until either a set number of expansion coefficients are generated or some energy threshold for the signal is reached. Each stage n yields an index to the dictionary structure denoted by γ_n, an expansion coefficient p_n, and a residual R_n which is passed on to the next stage. After M stages, the signal can be approximated by a linear function of the dictionary elements:

$$\hat{f}(t) = \sum_{n=1}^{M} p_n g_{\gamma_n}(t). \tag{4}$$

The above technique has useful signal representation properties. For example, the dictionary element chosen at each stage is the element which provides the greatest reduction in mean square error between the true signal $f(t)$ and the coded signal $\hat{f}(t)$. In this sense, the signal content is coded in order of importance, which is desirable in progressive transmission situations, or situations where the bit budget is limited. For image and video coding applications, the most visible features tend to be coded first, while weaker image features are coded later, if at all. The dictionary set we use in this paper consists of an overcomplete collection of 2-D separable Gabor functions [9].

At stage n of the algorithm, the largest inner product p_n, along with the indexing parameter γ_n define an atom. Indexing parameter γ_n consists of (a) the two parameters that define the 2-D separable Gabor functions, and (b) the two position parameters that define its location in the residual image. When the atom decomposition of a single residual frame is computed, it is important to code the resulting parameters efficiently to minimize the resulting bit rate. p_n and the parameters defining the structure of the dictionary are typically coded using fixed Huffman codes. In the remainder of this paper we investigate the coding of the position parameters.

3 Atom Position Coding

We begin this section by deriving an expression for the entropy of independent identically distributed (IID) position parameters of n atoms within a frame of size N pixels. We then provide a brief overview of existing position coding techniques.

3.1 Entropy for IID Data

Let us consider the problem of coding the position parameters of n atoms in a frame of size N pixels assuming the positions to be uniformly identically distributed within the frame and independent of each other. We allow several atoms to be at the same position, as it does happen in practice. It can be shown that there are $\binom{N+n-1}{n}$ possible placements of the n atoms within a frame with N pixels, and since the atom positions are IID, they are all equi-probable. Thus the entropy per atom is given by

$$\frac{1}{n}\log_2\binom{N+n-1}{n}. \qquad (5)$$

While in practice the atoms for coding the DFD for a video sequence are not IID, the above expression is still a useful reference for atom position coding. Indeed, achieving a lower average bits per atoms is a reasonable indicator of how well we are able to capture the true statistical structure of the atom positions.

3.2 Previous Position Coding Schemes

The position coding technique proposed by Neff and Zakhor [9] is frame based and is based on 2D run-length coding. A differential coding strategy employs three basic codeword tables. The first table P1 is used at the beginning of the frame line to indicate the horizontal distance from the left side of the image to the location of the first atom on the line. For additional atoms on the same line, the second table P2 is used to transmit inter-atom distances. The third table P3 indicates how many lines may be skipped before the next line containing coded atoms. We refer to this position coding scheme as P1P2P3.

A block-based approach was initially introduced in [2]. The method presented in [1] improves upon block-based position coding technique by increasing the coding efficiency without sacrificing error resilience, and works as follows. The atoms within each 16×16 block are first re-ordered according to a spiral scanning order and then coded differentially. Four different Variable Length Coding (VLC) tables are used to Huffman code the run-lengths depending on the number of atoms in it. We refer to this position coding scheme as MB. In [8], the block size is varied from frame to frame using rate optimization.

The NumberSplit algorithm for position coding introduced in [1], is based on divide-and-conquer. First, the total number of atoms T in a given frame is transmitted in the header. Then the image is divided into two halves along a larger dimension, and the number of atoms in the left or top half is coded and transmitted. The algorithm is applied recursively to the halves of the image until there are no more atoms in a given half image or until the size of the half image is one pixel. Note that if we assume that each atom falls uniformly and independently of other atoms onto either half, then the number of atoms in the first half is binomially distributed on $\{0, \cdots, T\}$ with probability of $\frac{1}{2}$. Since the atoms tend to cluster in the DFD where the motion prediction fails, the tails of the binomial distribution are emphasized according to a clustering parameter f. We then use the resulting distribution to build a Huffman table to code the number of atoms in the first half. Hence the NumberSplit algorithm exploits the spatial coherence of the atom positions. We refer to this position coding scheme as NS.

4 Our Proposed Position Coding Scheme

The aim of our proposed scheme is to take advantage of both spatial and temporal coherence of the atom positions. We know that the atoms tend to cluster in regions where the motion prediction fails. The atom positions are thus spatially highly non-stationary; as such, we take this into account in our coding scheme. Furthermore, from one frame to the next, these regions of high motion are correlated. If there is a large number of atoms in a given region of the DFD, it is likely that there was also a large number of atoms in the same region in the previous frame. In the case of a scalable video coder, a scheme for atom position coding that takes advantage of the temporal coherence was introduced in [5]. The principle is to represent the atom positions by a quadtree and to transmit the XOR of this quadtree with the quadtree of another group of atoms in the previous frame.

We propose a block-based position coding technique where the number of atoms in each block is coded differentially with respect to the same block in the previous frame. If the number of atoms in block b in frame n is $N_{b,n}$, we code the difference $N_{b,n} - N_{b,n-1}$. For the first P frame in the GOP, we code the number of atoms in each block, and for the subsequent P frames we code the difference of numbers of atoms. Thus, at the decoder, if frame $n-1$ has been correctly decoded we have access to $N_{b,n-1}$, and we can recover $N_{b,n}$. In doing so, we take advantage of the temporal coherence. The block size has to be chosen accordingly. Indeed, if the block size is too small, the correlation between $N_{b,n-1}$ and $N_{b,n}$ will be low and it is more efficient to code $N_{b,n}$ directly instead of $N_{b,n} - N_{b,n-1}$. Conversely, if the block size is too large, there are fewer blocks and we achieve lower gain since the number of symbols to code becomes small. We have found 16 × 16 blocks to be optimal for the video sequences we tested. We use an adaptive arithmetic encoder to code the symbols corresponding to $N_{b,n} - N_{b,n-1}$.

Once we have transmitted the number of atoms, we need to code the atom positions inside each of the 16×16 blocks. In our video coder the chroma channels are subsampled by a factor of 2 in the horizontal and vertical direction, and thus the positions of the UV atoms are subsampled by 2. In the existing schemes such as [1] the UV atom positions are doubled, and then all the YUV atoms are coded together. This is inefficient since two bits are lost for every UV atom. Our new scheme first subsamples the Y atom positions in the 16 × 16 block by two, so all the atoms inside that block are defined on an 8 × 8 grid. We code the Y, U and V atom positions using the NumberSplit algorithm with a clustering parameter set to 0.2 as applied to the 8 × 8 block. Then, for each atom, we send a codeword specifying if it is a Y, U or V atom, and for the Y atom we add two bits to specify its position in the 16 × 16 block. Since the atoms tend to lie at the boundaries of the blocks, the NumberSplit algorithm takes into account the non-uniformity of the atom positions inside the blocks. We refer to this position coding scheme as NS+TC.

5 Results

In this Section we present results to compare NS+TC with the frame-based coding technique (P1P2P3), the macro-block based position coding (MB), and the NumberSplit algorithm (NS) described earlier. We use the following QCIF training sequences to compute the frequency tables used to build Huffman tables and for the arithmetic encoder: Silent, Mother, Coast and News. The QCIF test sequences are Akiyo, Carphone, Claire, Container, Foreman, Grandma, Hall, Highway, Salesman and Sean. To approximate constant quality , we encode each frame such that the maximum mean square error (MSE) computed over each macro-block is smaller than a target. We use the values 5, 10 , 20 and 30 as target MSEs. We have found empirically that using values lower than 5 does not offer further visual improvements, and the resulting atoms mostly correspond to noise. The frame rate is 10 Hz, and we code one I frame followed by 99 P frames. We compare the effectiveness of each position coding technique by examining the number of position bits and its impact on the overall bit rate. We also examine the average number of bits per atom for each technique and compare it to the theoretical value, assuming the data is IID. The results are presented in Table 1 for the comparison of the average number of bits per atom position, and in Table 2 for the comparison of the overall bit rate using the different position coding techniques. Our proposed scheme NS+TC outperforms all the

Table 1. Comparison of the average number of bits per atom position

video sequence	MSE	atoms per frame	IID	P1P2P3	MB	NS	NS+TC
Akiyo	5	151	8.80	8.54	8.77	8.06	7.49
	10	86	9.59	9.54	9.57	9.09	8.43
	20	47	10.43	10.59	10.87	10.09	9.71
Carphone	5	490	7.11	8.08	8.82	7.01	6.51
	10	289	7.87	8.98	8.54	7.83	7.27
	20	165	8.67	9.90	9.10	8.73	8.12
Claire	5	161	8.71	8.47	8.78	7.86	7.04
	10	90	9.53	9.32	9.25	8.76	7.95
	20	49	10.37	10.32	10.58	9.85	9.4
Container	5	384	7.46	7.76	8.21	7.04	6.51
	10	207	8.35	8.50	8.70	7.84	7.18
	20	114	9.19	8.96	9.13	8.57	7.91
Foreman	10	345	7.62	8.78	8.59	7.75	7.30
	20	205	8.36	9.60	8.86	8.52	8.13
	30	148	8.83	10.14	9.20	9.02	8.47
Grandma	5	270	7.97	8.21	8.40	7.65	7.01
	10	141	8.89	9.22	8.98	8.73	7.90
	20	68	9.92	10.32	10.16	9.87	9.18
Hall	5	404	7.39	7.74	8.29	7.27	6.32
	10	185	8.51	8.61	8.76	7.95	7.18
	30	56	10.19	9.25	10.03	9.11	8.96
Highway	5	565	6.90	7.41	8.41	6.99	5.72
	10	217	8.28	8.67	8.54	7.92	6.85
	20	72	9.84	10.35	9.85	9.45	9.07
Salesman	5	300	7.82	8.31	8.43	7.54	7.07
	10	170	8.63	9.16	8.96	8.45	7.87
	20	93	9.48	9.92	9.75	9.34	8.78
Sean	5	238	8.15	8.10	8.59	7.47	7.09
	10	140	8.90	8.92	9.17	8.36	7.91
	20	82	9.66	9.81	9.90	9.23	8.84

other schemes in every case. The average number of bits to code the position of an atom is lower than the theoretical value in the case of IID data. For most of the sequences the reduction exceeds 1 bit per atom. This shows that our scheme effectively captures the spatial and temporal coherence. Furthermore, by comparing NS+TC to NS, our scheme uses an average of 0.6 fewer bits. Since NumberSplit exploits the spatial coherence only, this indicates that taking into account the temporal coherence results in a significant gain in coding the atom positions.

The two last columns of Table 2 show the reduction in bit rate when using NS+TC instead of MB, and the percentage of bits used to code the atom positions using MB. Since the percentage of bits used to code the atom positions can exceed 40% of the overall bit rate, there is a significant reduction in the

Table 2. Comparison of the overall bit rate in kbps

video sequence	MSE	atoms per frame	P1P2P3	MB	NS	NS+TC	bit rate gain NS+TC vs MB	MB position bits %
Akiyo	5	151	32.42	32.76	31.66	30.81	6.0%	40.3%
	10	86	20.40	20.41	20.00	19.44	4.8%	39.3%
	20	47	12.94	13.03	12.69	12.50	4.1%	38.5%
Carphone	5	490	108.94	114.86	103.47	101.14	11.9%	38.8%
	10	289	72.94	72.05	69.37	67.88	5.8%	33.8%
	20	165	49.02	47.81	46.90	45.99	3.8%	30.8%
Claire	5	161	36.27	36.83	35.27	33.97	7.8%	38.2%
	10	90	22.85	22.80	22.34	21.61	5.2%	36.5%
	20	49	14.62	14.72	14.38	14.14	3.9%	34.6%
Container	5	384	73.83	75.61	71.05	69.05	8.7%	40.1%
	10	207	42.22	42.63	40.86	39.50	7.3%	37.6%
	20	114	25.12	25.32	24.68	23.93	5.5%	37.6%
Foreman	10	345	90.11	90.00	86.55	85.12	5.4%	33.1%
	20	205	64.10	62.15	61.28	60.28	3.0%	26.8%
	30	148	52.50	51.23	50.84	50.10	2.2%	26.4%
Grandma	5	270	58.43	58.99	56.88	55.18	6.5%	38.4%
	10	141	32.29	31.94	31.58	30.43	4.7%	39.4%
	20	68	18.27	18.15	17.97	17.49	3.6%	37.5%
Hall	5	404	86.83	89.11	84.94	81.11	9.0%	37.6%
	10	185	41.87	42.18	40.65	39.25	6.9%	33.7%
	30	56	15.01	15.45	14.94	14.85	3.9%	33.5%
Highway	5	565	130.05	135.79	127.66	120.49	11.3%	34.7%
	10	217	54.15	53.88	52.51	50.18	6.9%	34.5%
	20	72	22.28	21.92	21.63	21.36	2.6%	32.1%
Salesman	5	300	64.3	64.82	61.96	60.64	6.4%	38.9%
	10	170	38.65	38.29	37.41	36.45	4.8%	39.3%
	20	93	24.25	24.05	23.69	23.16	3.7%	37.2%
Sean	5	238	49.87	51.10	48.18	47.32	7.4%	39.4%
	10	140	30.82	31.20	29.95	29.33	6.0%	40.4%
	20	82	20.16	20.15	19.68	19.31	4.2%	38.7%

overall bit rate when using a more efficient position coding technique such as NS+TC. We observe that the performance improves with the number of atoms to code. Indeed, at higher bit rates, the percentage of bits allocated to atom coding becomes higher, resulting in more atoms to be coded, and hence larger gain by using a more efficient position coding method. We observe this trend for all of the encoded sequences in Table 2. The lower the MSE target, the larger the reduction in bit rate. The average reduction for MSE values 5, 10 and 20 are 8.3%, 5.8% and 3.8% respectively. The largest gain is obtained with the sequence Highway coded with MSE 5 where the reduction in bit rate is 11.3%. The sequence Foreman has the lowest reduction in bit rate of all sequences, due to its high motion nature and significant percentage of bits going to motion.

6 Conclusion

In this paper we have proposed a new scheme to code the positions of the atoms in a Matching Pursuit based video encoder. This new position coding method exploits the statistical structure of the DFD, namely the spatial and temporal coherence. It results up to 11% lower bit rate as compared to existing schemes.

References

1. Al-Shaykh,O., Miloslavsky, E., Nomura, T., Neff, R., and Zakhor, A.: Video Compression using Matching Pursuits.IEEE Trans. on Circuits and Systems for Video Tech. **9** no.1 (1999) 123-14.
2. Banham, M. and Brailean, J.: A selective update approach to matching pursuits video coding. IEEE Trans. on Circuits and Systems for Video Tech. **7** no.1 (1997) 119-129.
3. Friedman, J. H., and Stuetzle, W.: Projection Pursuit Regression. Journal of the American Statistical Association **76**, (1981) 817-823.
4. Frossard, P., Vandergheynst, P., Figueras I Ventura, I. M., and Kunt, M.: A Posteriori Quantization of Progressive Matching Pursuit Streams. IEEE Trans. on Signal Processing **52**, no.2 (2004) 525-535.
5. Lin, J-L., Hwang, W-L., and Pei, S-C.: SNR Scalability Based on Bitplane Coding of Matching Pursuit Atom s at Low Bit Rates: Fine-Grained and Two-Layer. IEEE Trans. on Circuits and Systems for Video Tech. **15** no.1 (2005).
6. Mallat, S., and Zhang, Z.: Matching Pursuits with time-frequency dictionaries. IEEE Trans. on Signal Processing **41**, No.12 (1993) 3397-3415.
7. Monro, D.M., and Poh, W.: Improved Coding of Atoms in Matching Pursuits. ICIP 2003, Barcelona (2003).
8. Moschetti, F., Sugimoto, K., Kato, S., and Etoh, M.: Bidimensional Dictionary and Coding Scheme for a Very Low Bitrate Matching Pursuit Video Coder. ICIP 2004, Singapore (2004).
9. Neff, R., and Zakhor, A.: Very Low Bit-Rate Video Coding based on Matching Pursuits. IEEE Trans. on Circuits and Systems for Video Tech. **7** no.1 (1997) 158-171.

Efficient Digital Pre-filtering for Least-Squares Linear Approximation

Marco Dalai, Riccardo Leonardi, and Pierangelo Migliorati

University of Brescia, Via Branze 38, Brescia 25123, Italy
{name.surname}@ing.unibs.it

Abstract. In this paper we propose a very simple FIR pre-filter based method for near optimal least-squares linear approximation of discrete time signals. A digital pre-processing filter, which we demonstrate to be near-optimal, is applied to the signal before performing the usual linear interpolation. This leads to a non interpolating reconstruction of the signal, with good reconstruction quality and very limited computational cost. The basic formalism adopted to design the pre-filter has been derived from the framework introduced by Blu et Unser in [1]. To demonstrate the usability and the effectiveness of the approach, the proposed method has been applied to the problem of natural image resampling, which is typically applied when the image undergoes successive rotations. The performance obtained are very interesting, and the required computational effort is extremely low.

1 Introduction

Linear interpolation is one of the simplest methods for digital to analog conversion, and it is still applied in applications where the conversion computational cost must be kept under control. Nevertheless, linear interpolation introduces some artifacts, e.g., blurring in image processing, and it is often necessary to use higher degree polynomials to achieve an acceptable level of approximation.

An interesting point is to establish if a signal needs to be actually interpolated or whether a non interpolating least-square approximation could be preferable. In literature the choice seems to have been interpolation, even when it does not represent a requirement and it clearly generates analog reconstructed signals with a much higher mean square error. A significant example is represented by the problem of image resampling when the new pixels are uniformly distributed with respect to the old ones, as in the case of image rotations by angles which are not multiples of $\pi/2$.

The concept of least-squares reconstruction of signals has been well studied in modern sampling theory (e.g., see [2]), and the solution is known to be given by an analog pre-filtering of the signal before sampling. The optimal filter inpulse response depends on the generating function used in the reconstruction step (and is called *dual* of the generating function itself).

In this paper we propose the introduction of a very simple digital filter to be applied to the signal before performing the classical linear interpolation, as an

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 161–169, 2006.
© Springer-Verlag Berlin Heidelberg 2006

equivalent of the analog analysis filter which is missing in our implementation. The approach is a direct extension of the results described by Blu and Unser in [1]. We show here that a very simple FIR filter leads to almost the same performance with respect to an optimal analog analysis pre-filter under the hypothesis that the signal has been sampled according to Shannon's sampling theorem. In this paper we will use a normalized frequency representation, so that this hypothesis is equivalent to consider the signal to be bandlimited to frequencies $|f| < 1/2$.

The rest of the paper is organized as follows. Section 2 describes some known results of interpolation and sampling theory, whereas the proposed digital pre-filter is introduced and discussed in Section 3. Experimental results showing the effectiveness of the proposed approach are reported in Section 4. Finally, concluding remarks are drawn in Section 5.

2 Interpolation and Sampling Theory

Given a set of values $s(k)$, corresponding to samples of a signal s at integer points, a linear interpolation constructs a piecewise linear approximation $\tilde{s}$ given, for $k \in N$ and $\delta \in [0, 1[$, by

$$\tilde{s}(k+\delta) = s(k)(1-\delta) + s(k+1)\delta$$

This is known to be equivalent, for $t \in \mathbb{R}$, to

$$\tilde{s}(t) = \sum_{k\in\mathbb{Z}} s(k)\, \Lambda(t-k) \tag{1}$$

where $\Lambda(t)$ is the unity triangular pulse defined by $\Lambda(t) = \max(0, 1-|t|)$.

Equation (1) is a special case of a general model used in modern sampling theory: given a reconstruction function φ a signal approximation can be obtained using the linear expansion

$$\tilde{s}(t) = \sum_{k\in\mathbb{Z}} c_k\, \varphi(t-k) \tag{2}$$

where the coefficients c_k have to be set depending on φ, s, and the desired characteristics of the approximation result. A very useful choice is to construct an *analysis* function $\tilde{\varphi}$, dependent on φ, and consider coefficients c_k of the form

$$c_k = \langle s, \tilde{\varphi}_k \rangle$$

where $\tilde{\varphi}_k(t) = \tilde{\varphi}(t-k)$. This assumption corresponds to filtering s with a filter with impulse response $h(t) = \tilde{\varphi}(-t)$, followed by an ideal sampler to obtain the sequence c_k.

With this formalism in mind, given a certain reconstruction function φ, the approximation $\tilde{s}$ is simply determined by the choice of $\tilde{\varphi}$. If we want to obtain the least-square approximation solution (in the form of eq. (2)), the analysis function is given by (see [2])[1]

$$\hat{\tilde{\varphi}} = \frac{\hat{\varphi}(f)}{\hat{a}_{\varphi}(f)}$$

where $a_{\varphi}(t)$ is the sampled autocorrelation of the function φ, and thus satisfies

$$\hat{a}_{\varphi}(f) = \sum_{k \in Z} |\hat{\varphi}(f-k)|^2$$

In this case $\tilde{\varphi}$ is called *dual* of φ and will be denoted with $\overset{\circ}{\varphi}$.

If we are interested to linear approximation, the expression of $\overset{\circ}{\varphi}\ (f)$ can be easily derived, given that $\varphi(t) = \Lambda(t)$ and $\hat{\varphi}(f) = \mathrm{sinc}^2(f)$. In this case one has

$$\hat{a}_{\Lambda}(f) = \frac{2}{3} + \frac{1}{3}\cos(2\pi f) \tag{3}$$

and, thus,

$$\hat{\overset{\circ}{\Lambda}} = \frac{3\,\mathrm{sinc}(f)^2}{2 + \cos(2\pi f)}$$

Consider that this function is even and real valued, then $\Lambda(t)$ must be even as well. Thus we have $h(t) =\overset{\circ}{\Lambda}\ (-t) =\overset{\circ}{\Lambda}\ (t)$, and we will refer to the dual function $\overset{\circ}{\Lambda}$ as the analysis filter.

3 The Proposed Digital Pre-filter

Once we have the expression of the analog analysis filter for the least-squares approximation this should be applied to the signal before sampling, in view of a future piecewise linear reconstruction (which is optimal according to the least squares criterion).

Obviously, this cannot be implemented in practice as the signal is normally available already in a digital form. So, the overall processing must be implemented in an equivalent way in the digital domain. Assuming that the sampling rate satisfies Shannon's sampling theorem, a near equivalent FIR pre-filter can be derived and applied to the digital signal providing nearly the same performance as its optimal analog counterpart.

In this way, it will be possible to reconstruct a piecewise linear least-square approximation of the original signal in a very efficient manner.

[1] $\hat{\varphi}$ represents the Fourier transform of the function φ. All functions with the superscript ^ will in what follows always identify the Fourier transform of its argument, unless it is explicitely stated differently in the text.

3.1 Filter

We choose to limit the number of taps of the near equivalent FIR filter to 5. This could be considered a completely arbitrary choice, but it will be shown that the quality of the approximation is practically unaffected by such limitation. In addition, this allows to maintain a very low computational cost. Longer filters have been considered without obtaining a substantial reduction in the mean square error.

As previously mentioned, the FIR digital filter should behave similarly to its Analog counterpart $\overset{\circ}{\Lambda}$. The design approach basically equates the first terms of the Taylor series expansion of both frequency representation of the two filters around $f = 0$. Given the symmetries of $\overset{\circ}{\Lambda}$, $h[n]$ has to be even shaped and satisfy a unity gain at $f = 0$. Thus its generic expression in the Z transform domain is given by

$$h(z) = 1 - a - b + \frac{a}{2}(z^1 + z^{-1}) + \frac{b}{2}(z^2 + z^{-2})$$

The associated Discrete Time Fourier Transform is thus

$$\hat{h}(f) = 1 - a - b + a\cos(2\pi f) + b\cos(4\pi f)$$

Using now its Taylor series expansion and limiting it to the fourth order approximation, one can rewrite

$$\hat{h}(f) = 1 + (2a - 8b)\pi^2 f^2 + \frac{32b - 2a}{3}\pi^4 f^4 + o(f^5)$$

Similarly the Taylor Series representation of the dual function is given by

$$\hat{\overset{\circ}{\Lambda}}(f) = 1 + \frac{1}{3}\pi^2 f^2 + \frac{2}{45}\pi^4 f^4 + o(f^5)$$

Equating the coefficients of f^2 and f^4 we obtain a linear system with two unknown, the solution of which leads to $a = -11/45$ and $b = 7/360$. Thus the filter transfer function is given by

$$h(z) = \frac{49}{40} - \frac{11}{90}(z^1 + z^{-1}) + \frac{7}{720}(z^2 + z^{-2})$$

The joint concatenation of this digital filter with a linear interpolator can also be viewed as an approximation process in the form of equation (2). The coefficients c_k correspond to the samples $s(k)$, and the reconstruction function φ is given by a linear combination of $\Lambda(t)$, $\Lambda(t \pm 1)$ and $\Lambda(t \pm 2)$ leading to the impulse response shown in fig 1.

3.2 Approximation Error

Equipped with the analytical expression of our filter, it is important to compare its performance with that obtained by the optimal solution. In [3] and [1], the authors gave extremely useful mathematical tools for the study of the mean

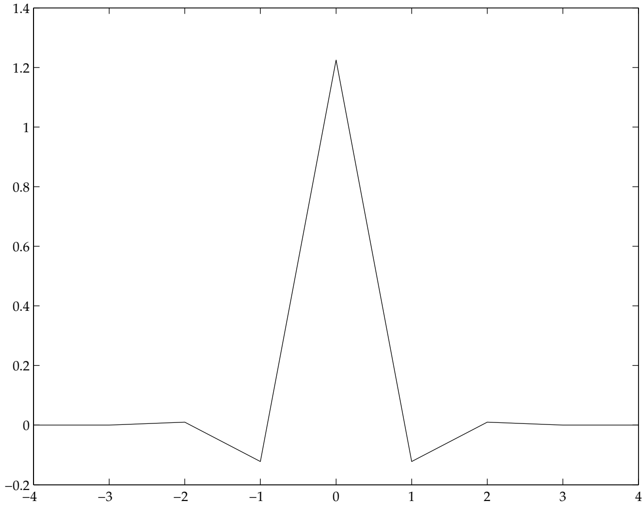


Fig. 1. Equivalent reconstruction function of the FIR pre-filtering method

square error behavior in approximation techniques. In particular it is shown that, if T is the sampling rate, a very good estimation of the mean square approximation error is given by

$$\eta_s(T) = \left(\int_{-\infty}^{\infty} |\hat{s}(f)|^2 E_{\tilde{\varphi},\varphi}(fT) df \right)^{1/2}$$

where the $E_{\tilde{\varphi},\varphi}(f)$ is called the *error kernel* function and is given by

$$E_{\tilde{\varphi},\varphi}(f) = |1 - \hat{\tilde{\varphi}}(f)^* \hat{\varphi}(f)|^2 + |\hat{\tilde{\varphi}}(f)|^2 \sum_{n \neq 0} |\hat{\varphi}(f+n)|^2 \tag{4}$$

or equivalently

$$E_{\tilde{\varphi},\varphi}(f) = 1 - \frac{|\hat{\varphi}(f)|^2}{\hat{a}_{\varphi}(f)} + \hat{a}_{\varphi}(f) |\hat{\tilde{\varphi}}(f) - \hat{\overset{\circ}{\varphi}}(f)|^2 \tag{5}$$

For our purpose we set T to 1, as we are using a normalized frequency representation. We have therefore to compare the error kernel for the least square approximation using piecewise linear functions, the classic linear interpolation, and our near optimal solution.

For the first case, we have that $\tilde{\varphi} = \overset{\circ}{\Lambda}$ and, calling $E_{\min} = E_{\overset{\circ}{\Lambda},\Lambda}$, from (5) and (3),

$$E_{\min}(f) = 1 - \frac{3\,\mathrm{sinc}(f)^4}{2 + \cos(2\pi f)}$$

For the classic linear interpolation, instead, $\tilde{\varphi}(t) = \delta(t)$ and thus, from (4),

$$E_{\delta,\Lambda}(f) = \frac{5}{3} + \frac{1}{3}\cos(2\pi f) - 2\,\mathrm{sinc}^2(f)$$

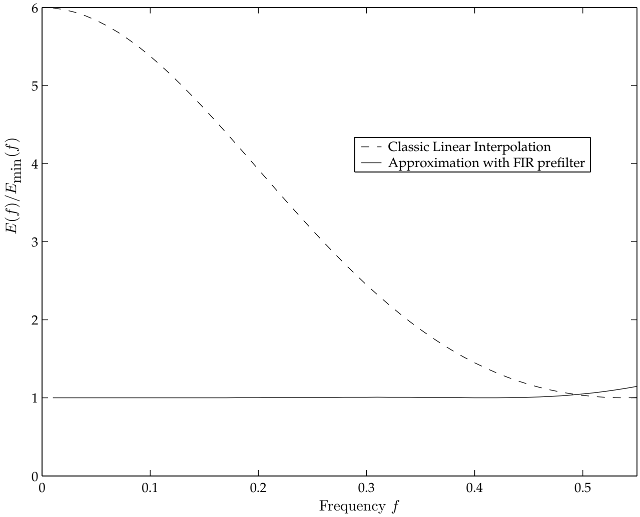


Fig. 2. Error kernels comparison

Finally, for the proposed FIR prefilter, $\tilde{\varphi} = h$ and, from (4), we obtain

$$E_{h,\Lambda}(f) = 1 + h^2(f)\left(\frac{2}{3} + \frac{1}{3}\cos(2\pi f)\right) - 2h(f)\mathrm{sinc}^2(f)$$

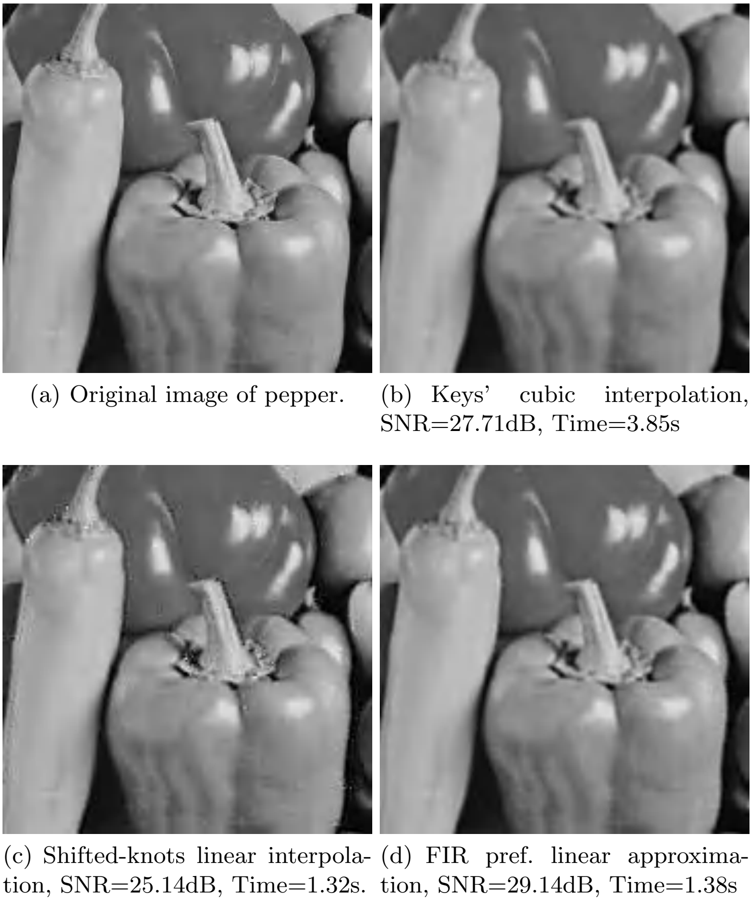
Comparing the power series expansion of these three functions, one may verify that $E_{\delta,\Lambda}(f) - E_{\min}(f) = O(f^4)$ while $E_{h,\Lambda}(f) - E_{\min}(f) = O(f^{12})$ which means that for low frequencies the FIR pre-filter leads to much smaller errors with respect to the classical interpolation. In Fig. 2 the normalized error kernels $E(f)/E_{\min}(f)$ are plotted for both methods. It clearly indicates that the proposed FIR pre-filtering method is practically identical to that of the least square solution up to the Nyquist frequency $f = 1/2$. Therefore, if the signal can be considered bandlimited to $|f| < 1/2$ (i.e., Shannon's sampling theorem is satisfied), the solution obtained by FIR pre-filtering can be considered equivalent to the optimal analog solution.

4 Simulation Results

In order to demonstrate the efficiency of our approximation scheme, we compare it with two meaningful interpolation methods, namely Keys' cubic interpolation, which is the reference for high-quality image reconstruction methods, and the more recent interesting shifted-knots linear interpolation proposed by Blu *et al.* in [4]. We perform an image resampling test, based on successive rotation of 256 gray level images, that is a variant of that used in [4].

We randomly choose 14 angles θ_i with $\theta_i \in [-\pi/2, \pi/2]$ for $1 \leq i \leq 14$ and we set $\theta_{15} = 0$. The value θ_i represents the angular position (with respect to the original one) of the image at the i-th step. The image at step $i+1$ is obtained by rotating the image at step i by an angle $\theta_{i+1} - \theta_i$. It is clear that at the 15-th step the image is exactly in the initial position and we can compare the rotated image with respect to the original one to evaluate the error introduced by successive resamplings of the rotated rectangular image. All computations have been performed with floating point arithmetic so as to avoid the effects of quantization. Furthermore, the error has been computed only on the central part of the image so that boundaries effect have been discarded.

This simulation experiment is different from that proposed in [4] for which the image was rotated 15 times by an angle $2\pi/15$ in the same direction. Such a choice, in fact, causes a compensation of the phase distortion potentially introduced (and it is the case for the shifted-knots method) by the approximation system, due to the presence of near-opposite angles during the simulation. This



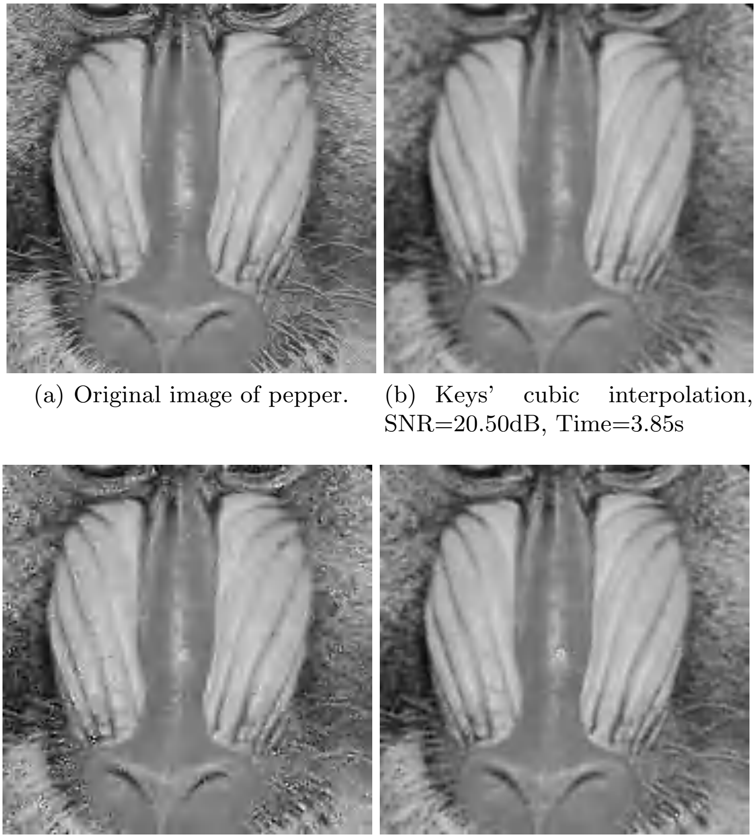
(a) Original image of pepper.

(b) Keys' cubic interpolation, SNR=27.71dB, Time=3.85s

(c) Shifted-knots linear interpolation, SNR=25.14dB, Time=1.32s.

(d) FIR pref. linear approximation, SNR=29.14dB, Time=1.38s

Fig. 3. Results of simulations on the image Pepper



(a) Original image of pepper.

(b) Keys' cubic interpolation, SNR=20.50dB, Time=3.85s

(c) Shifted-knots linear interpolation, SNR=18.02dB, Time=1.32s

(d) FIR pref. linear approximation, SNR=21.42dB, Time=1.38s

Fig. 4. Results of simulations on the image Pepper

is the reason for which, here, shifted-knots linear interpolation gives smaller SNR values than Keys' cubic one (even if the images can be considered visually more satisfactory), contrarily to what reported in [4].

Figures 3 and 4 show the results of the test for the Pepper and Baboon images, respectively. It is important to note that the FIR-prefilter linear approximation can give higher quality (both visual and with respect to the SNR value) than Keys' cubic interpolation in 1/3 of the computation time.

Furthermore, it is interesting to compare the performance of our method to that of the shifted-knots one; the phase distortion of the latter, in fact, leads to a lower SNR value in our rotation test. Visually, the obtained images are not blurred (as with classic linear interpolation) but they are affected by the presence of shot-noise. The least-squares linear approximation, on the contrary, does not introduce this artifact and gives images with very little blur, leading to higher quality for about the same computational cost.

5 Conclusion

In this paper we have presented the idea of using a very simple linear phase FIR pre-filter to compute a least square linear approximation of a digital signal for bandlimited analog signals reconstruction. The digital pre-processing filter is applied to the signal before performing the classic linear interpolation, so as to obtain a non interpolating analog approximation. The usefulness of the approach has been demonstrated by applying it to the problem of resampling rotated images. In this context, the need for an exact interpolation of the original pixels is not a requirement, and the quality of the rotated signal remains very close to the original.

References

1. Blu, T., Unser, M.: Quantitative Fourier analysis of approximation techniques: part I–interpolators and projectors. IEEE Trans. Signal Process. **47**(10) (1999) 2783–2795
2. Unser, M.: Sampling-50 years after Shannon. Proc. IEEE **88**(4) (2000) 569–587
3. Blu, T., Unser, M.: Approximation error for quasi-interpolators and (multi)-wavelet expansions. Appl. Comput. Harmon. Anal. **6**(2) (1999) 219–251
4. Blu, T., Thvenaz, P., Unser, M.: How a simple shift can significantly improve the performance of linear interpolation. Proc. IEEE Int'l Conf. on Image Proc. (2000) III.377–III.380

Motion Compensated Frame Rate Up-Conversion for Low Frame Rate Video

Kenji Sugiyama[1], Mitsuhiro Fujita[2], Takahiro Yoshida[2], and Seiichiro Hangai[2]

[1] Seikei University, Faculty of Science and Technology,
3-3-1 Kichijoji-kitamachi Musashino-shi, Tokyo 180-8633, Japan
sugiyama@st.seikei.ac.jp
[2] Tokyo University of Science, Faculty of Engineering,
1-3 Kagurazaka, Shinjuku, Tokyo 162-8601, Japan
hangai@ee.kagu.tus.ac.jp

Abstract. Frame rate up-conversion is useful for low frame rate pictures such as movie. In case of conversion, basic motion compensation causes gap at covered area and uncovered area. To avoid the gap, we propose the motion compensation centered at the position of target. Fully motion compensated interpolation is possible in this method. To realize this, error normalizing by block activity is used for motion estimation. We also propose the adaptive interpolation by using four frames. Finally, we compare the original picture with the interpolated picture converted from decimated picture. The proposed conversion method gives 5dB better at the maximum PSNR than the conventional method.

1 Introduction

Frame rate up-conversion technology is useful for low frame rate pictures such as film movie or mobile video system. In liquid crystal display, long period lighting causes a degradation of motion picture quality. Up-conversion of frame rate solves this problem.

In case of conversion, frames should be made by motion compensated (MC) inter-picture processing. However, basic MC for coding causes gap at covered and uncovered area[1-2].

To avoid the gap, Bi-directional interpolation is useful[3]. We propose the MC centered at the position of target. Fully MC interpolation is possible in this method. To realize this, we discuss some method of motion estimation and adaptive interpolation by using four frames.

To evaluate proposed method, we use 60fps picture. 30fps pictures decimated from 60fps are converted to 60fps. The converted 60fps pictures are compared to the original pictures. PSNR and subjective picture quality are checked.

2 Purpose

Figure 1 shows application systems of frame rate up-conversion. Frame rate of movie film is 24fps (frame per second). Frame rate of picture phone or mobile broadcast is

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 170 – 178, 2006.
© Springer-Verlag Berlin Heidelberg 2006

usually 15fps. To realize smooth motion picture in such system, frame rate up-conversion after decoding is necessary. In this figure “60i” means interlaced scanning pictures with 60 fields per second.

LCD has disadvantage of motion picture quality because of its long lighting period. This causes blur (de-focusing effect) in high motion picture. This problem is also solved by frame rate up-conversion.

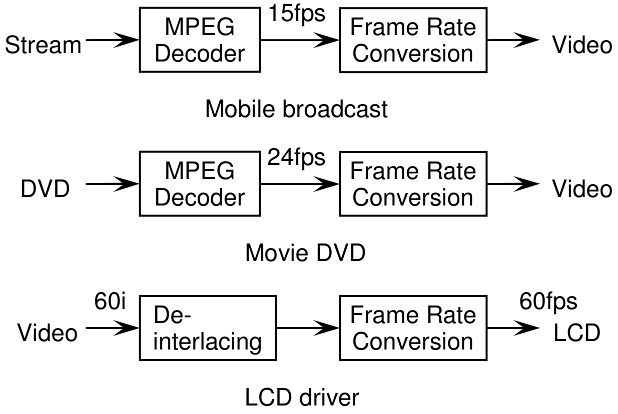


Fig. 1. Application systems of frame rate conversion

3 Frame Rate Conversion

Figure 2 shows a up-conversion of picture rate in the ratio of two, such as 30fps to 60fps. The frame in every other frame is made from the existing frame of order by interpolation.

To realize frame rate up-conversion, motion compensation (MC) at frame interpolation is necessary. In case of Non-MC, duplication of frame does not improve smoothness, and, averaging of frames causes overlapped picture. However, MC at interpolation is not easy.

In a MC of coding, a target frame and reference frames exist really. However, in case of picture rate conversion, target is at virtual position. This causes some difficulties in

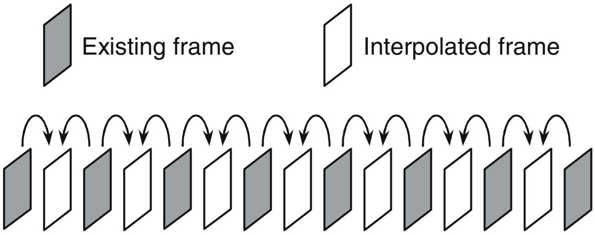


Fig. 2. Frame rate conversion

motion estimation (ME) and MC. Simple motion vector (MV) of picture coding generates indefinite area (gap) in figure 3 left. This gap is filled by meaningless non-MC picture usually. We should solve this problem.

3.1 MC Centered at Target

To avoid the gap, we propose new MC that centered at target frame position. In this method, block of MC is divided at target frame and MC is applicable fully. Figure 3 right side shows example of motion vector in this method.

Equation 1 is a conventional MC, and, Equation 2 is the proposed MC. In case of Proposed MC, MC is applicable for all pixels.

$$P_t(x+dx, y+dy) = P_0(x+2dx, y+2dy) \quad (1)$$

$$P_t(x, y) = \{P_0(x+dx, y+dy)+ P_1(x-dx, y-dy)\}/2 \quad (2)$$

where

P_t: pixel of target frame, P_0: pre frame, P_1: post frame

x, y: pixel position dx,dy: motion vector (MV)

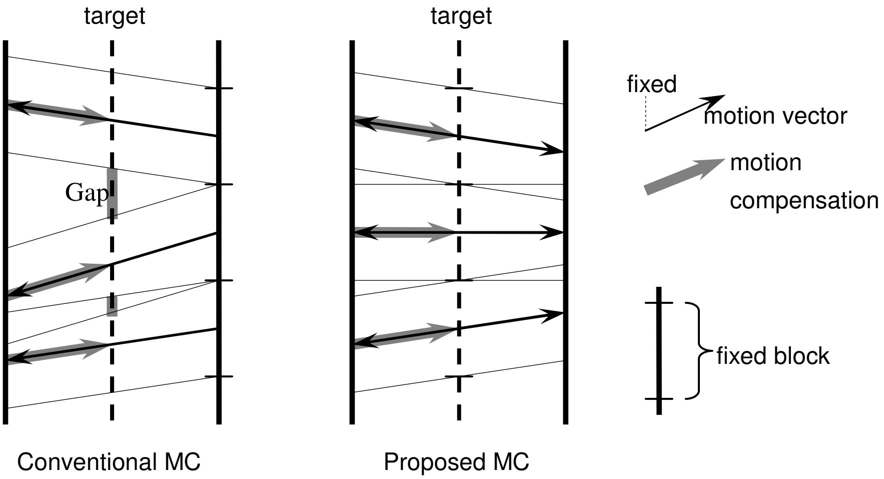


Fig. 3. MV relation on MC

3.2 Motion Estimation Using Normalized SAD

Summation of absolute difference (SAD) between target frame and reference frame is useful to decide MV in ME usually. However, in case of proposed MC, both reference frame moves in ME. This means that minimum SAD shown by equation 3 can not give right MV.

$$SAD(x,y)=\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}|P_0(x+dx+i,y+dy+j)-P_1(x-dx+i,y-dy+j)| \quad (3)$$

where N,M: block size.

To solve this problem, SAD should be changed to comparable value at any spatial position. SAD very depends on spatial activity. Consequently, we should normalize SAD by spatial activity. A block activity ACT(x,y) as spatial activity is estimated and used for normalization of SAD. MV is decided by the normalized SAD shown by equation 4.

$$SAD_{NORM}(x, y) = SAD(x, y) / ACT(x, y) \tag{4}$$

Considering of processing complexity and performance, we take two pixel differences for the block activities.

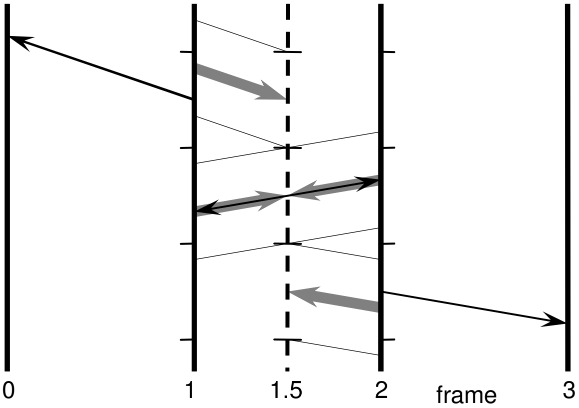


Fig. 4. Adaptive interpolation using four frame

3.3 Adaptive Interpolation Using Four Frames

Another problem of interpolation is detail change by object motion. It is caused at covered area and uncovered area. To solve this problem, we apply adaptive interpolation with three interpolations mode (forward, backward or bi-directional).

Basically, we use bi-directional interpolation. In case of usual picture area, it adopts to small change of picture. Noise is suppressed in this interpolation. But, bi-directional interpolation is not useable at covered area and uncovered area.

To interpolate such area, one directional MC (forward or backward) is useful. However, we can not estimate the MV of one directional MC by using frame 1 and 2 in such area. We estimate the motion between frame 0 and 1 for forward MC, and frame 2 and 3 for backward MC.

Figure 4 shows example of adaptive interpolation. MV of bi-directional interpolation comes between frame 1 and 2. However, MV of one-directional interpolation comes from extended time between frame 0 and 1 (2 and 3).

4 Processing

To realize adaptive interpolation using four frames, we take four steps of processing as shown in Figure 5. At first, usual block matching gives SAD at all MV. Forward and backward MV is decided by SAD at the block which extended from target frame.

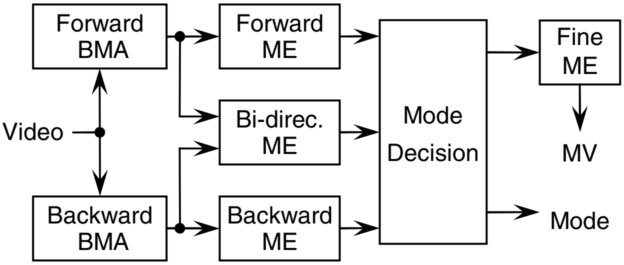


Fig. 5. Block diagram of ME

4.1 Block Matching at Both Directions

At first, usual block matching architecture (BMA) is used. In pair frames, One is fixed block another is searched. BMA is calculated at both relations. The forward and the backward are the same as a picture coding. Results of BMA are held for while, and, MV is not decided yet.

In consideration of block size and MV reliability, MC of 16x16 pixel is too large to represent object, MV of 8x8 pixel is less reliability. Then, we take 8x8 pixel block using weighted SAD of neighbor blocks.

4.2 MV Decision

One-directional MV (Forward, Backward) are simply calculated by SAD of block which MV reached. In this case, SAD between different block are compared. Normalizing is useful for decision by minimum SAD.

Bi-directional MV is calculated by using re-ME which centered on target frame. If MV of Forward is similar to MV of Backward (less than one pixel), initial MV of re-ME

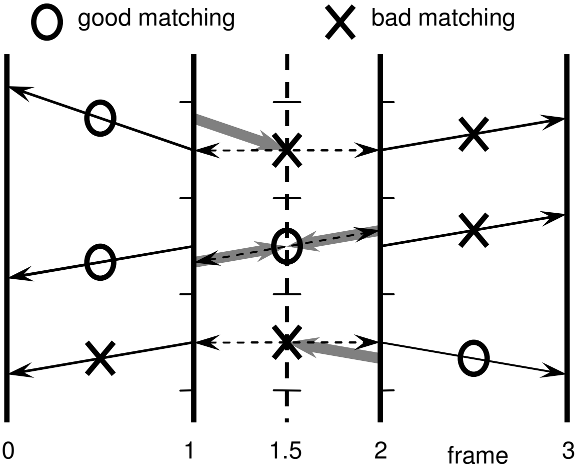


Fig. 6. Interpolation mode decision

is average of both MV. Search range is one pixel. If not, initial MV is zero (dx=0, dy=0). Search range is significantly wider. Search accuracy is one pixel in any case.

4.3 Mode Decision

Interpolation mode is decided based on three SAD_{Norm} between frame 0 and frame 1 ($D_{0\text{-}1}$), 1 and 2 ($D_{1\text{-}2}$), 2 and 3 ($D_{2\text{-}3}$). At first, we select One-directional or Bi-directional. If One-directional is selected, Forward or Backward is selected next. Figure 6 shows frame number and selection of the interpolation mode.

In usual motion area, Bi-directional will be good interpolation. Usual motion gives at least two small SAD_{Norm} include $D_{1\text{-}2}$.

To realize this, we compare three SAD_{Norm} to pre-fixed threshold value. In the other case, motion area is un-usual such as covered area or un-covered area.

Selection of Forward or Backward is simple. SAD_{Norm} of $D_{0\text{-}1}$ and $D_{2\text{-}3}$ are compared. If $D_{0\text{-}1}$ is less than $D_{2\text{-}3}$, Forward is used. In another case, Backward is used.

4.4 Improvement of Motion Accuracy

MV is basically one pixel accuracy until interpolation mode selection. To improve accuracy of MC, we take re-ME which search around the MV of decided interpolation mode. It is popular processing of ME of coding. Finally, we take half pixel accuracy for all MV.

5 Experiments

To evaluate the performance of proposed frame rate conversion, we use picture quality check system shown in figure 7. We use 60fps progressive scanning (60p) picture. The 30p, which is decimated from 60p, is converted to the 60p. Converted 60p is compared to original 60p picture.

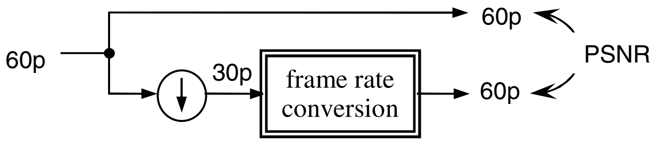


Fig. 7. Measurement processing

Table 1. ME/MC Parameters

Item	Conventional	Proposed	
Direction	One	One	Two(Bi)
Search range	+/-30 pixel	+/-30 pixel	+/-7 pixel
MC accuracy	0.5 pixel	0.5 pixel	0.5 pixel
Block size	8x8	8x8	8x8

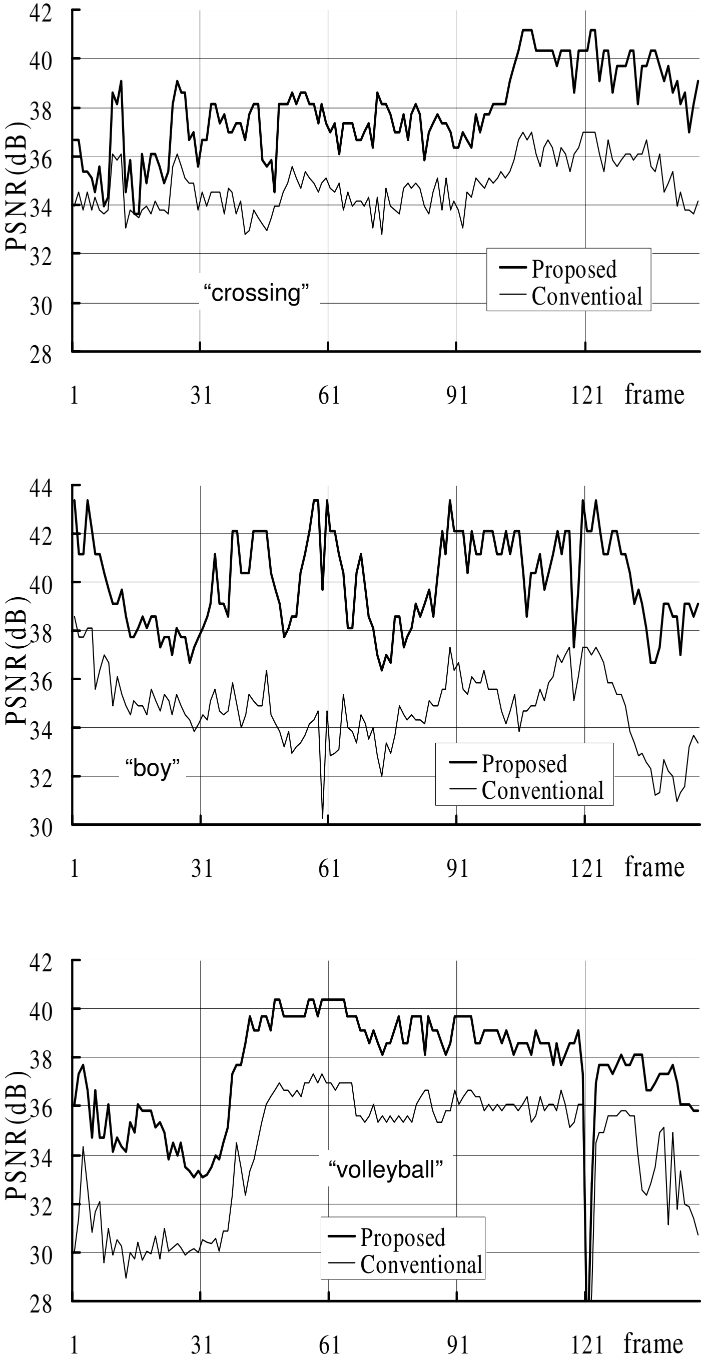
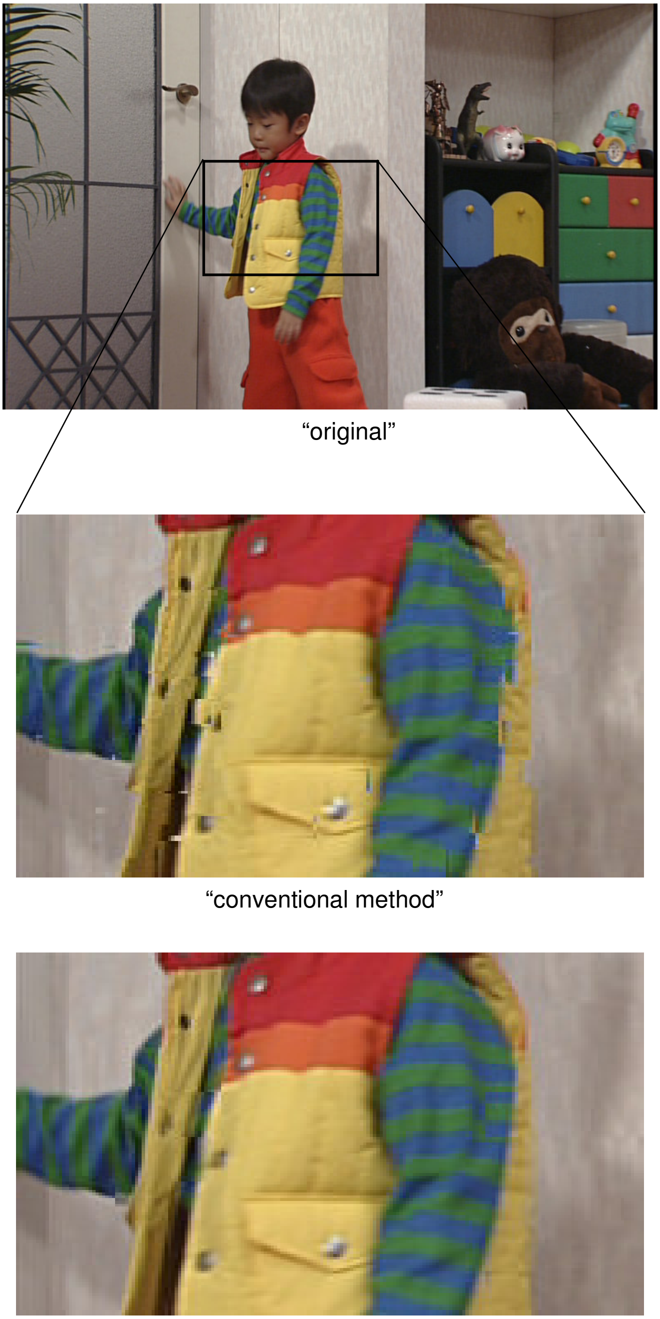


Fig. 8. PSNR of interpolated frames



“original”

“conventional method”

“proposed method”

Fig. 9. Example of interpolated frame “boy”

Processing parameters are shown in table1. Parameter of One-directional is the same as conventional method. In conventional method, “gap” is filled by Non-MC pixel.

Figure 8 shows experimental results of three sequences. “crossing” is stable camera picture including moving car and two girls. “boy” is moving camera picture with walking boy. "volleyball” is zooming picture at volleyball game. The proposed conversion method gives 5dB better at the maximum PSNR than the conventional method.

Figure 9 is example of converted picture. Proposed method shows significantly better picture quality at complex motion scene.

6 Conclusion

We propose a new frame rate conversion method using motion compensation centered at the position of target. Fully motion compensated interpolation is possible in this method. To realize this, activity normalizing is used for motion estimation. We also propose the adaptive interpolation by using four frames.

By comparing the original picture with the interpolated picture converted from decimated picture. It is found that the proposed conversion method gives 5dB better at the maximum PSNR than the conventional method. Subjective picture quality at complex motion scene is significantly improved. But, incorrect motion vector and mode selection still exist. We should improve the correctness of them.

References

1. Sung, W. R., Kang, E. K., and Choi, J. S.: Adaptive Motion Estimation Technique for Motion Compensated Interframe Interpolation. IEEE Trans. on Consumer Electronics **45** No.3 (1999) 753-761.
2. Kim, D. W., Kim, J. T., and Ra, I. H.: A New Video Interpolation Technique Based on Motion Adaptive Subsampling. IEEE Trans. on Consumer Electronics **45** No.3 (1999) 782-787.
3. Choi, B. T., Lee, S. H., and Ko, S. J.: New Frame Rare Up-Conversion Using Bi-directional Motion Estimation. IEEE Trans. on Consumer Electronics **46** No.3 (2000) 603-609, .
4. Ojo, O. A., and Schoemaker, H.: Adaptive Global Concealment of Video Up-Conversion Artifacts. IEEE Trans. on Consumer Electronics **47** No.1 (2001) 40-46.
5. Ha, T., Lee, S., and Kim, J.: Motion Compensated Frame Interpolation by new Block-based Motion Estimation Algorithm. IEEE Trans. on Consumer Electronics **50** No.2 (2004) 752-759.
6. Sugiyama, K., Aoki, T., and Hangai, S.: A Frame Rate Conversion Using Fully Motion Compensated Interpolation. Proc. of IEEE 2005 International Conference on Consumer Electronics, No.2.4.5 (2005).

Progressive Contour Coding in the Wavelet Domain

Nicola Adami, Pietro Gallina, Riccardo Leonardi, and Alberto Signoroni

Università degli Studi di Brescia, Telecommunications Group,
Dipartimento di Elettronica per l'Automazione,
via Branze, 38, Brescia I25123, Italy
{firstname.lastname}@ing.unibs.it
http://www.ing.unibs.it/tlc

Abstract. This paper presents a new wavelet-based image contour coding technique, suitable for representing either shapes or generic contour maps. Starting from a contour map (e.g. a segmentation map or the result of an edge detector process), a unique one-dimensional signal is generated from the set of contour points. Coordinate jumps between contour extremities when under a tolerance threshold represent signal discontinuities but they can still be compactly coded in the wavelet domain. Exceeding threshold discontinuities are coded as side information. This side information and the amount of remaining discontinuity are minimized by an optimized contour segment sequencing. The obtained 1D signal is decomposed and coded in the wavelet domain by using a 1D extension of the SPIHT algorithm. The described technique can efficiently code any kind of 2D contour map, from one to many unconnected contour segments. It guarantees a fully embedded progressive coding, state-of-art coding performance, good approximation capabilities for both open and closed contours, and graceful visual degradation at low bit-rates.

1 Introduction

From a human observer point of view visual content can often be captured by just relying on image discontinuities, often referred to as edges, contours, shapes,... depending on the application context. An efficient and functional coding of contour information, inherent to or extracted from images or video frames, can be exploited in order to improve the content analysis and management capabilities of picture and video archiving or the effectiveness of visual communication systems. There is a wide literature regarding the so called "shape coding" case (i.e. the coding of one or more closed contour lines that typically correspond to the boundaries of segmented objects) because of its role in object based video coding approaches. However, as far as we know, there is no proposed method for effective lossy coding of generic contour maps.

Currently, the best performing techniques work in the data domain and can be subdivided into two families: line-based methods [2, 3] and bitmap-based methods [4, 5].

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 179–188, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Transform coding of contours has already been introduced with the Fourier descriptors method [1], which is however suited only for connected and closed contour lines. Moreover, their coding performance are not competitive with respect to the state-of-art of shape coding. Instead of what has happened for image coding, the wavelet transform has not yet been considered in any kind of shape or contour map coding. In this work we set up the generic contour map coding problem as a one-dimensional (1D) signal coding problem. The 1D signal to code is directly obtained from the sequence of coordinates of the contour segments that define the contour map. The 1D signal generation is not unique because it depends on the contour segments concatenating order. We will show how to optimize the representation of this structural information so as to build a low complexity 1D signal and a limited amount of side information. The wavelet representation of the so obtained 1D signal guarantees an efficient coding of the contour map itself and allows to exploit the typical wavelet transform coding features and bit-stream properties (such as progressive quality reconstruction and, presumably, spatial scalability).

The paper is organized as follows: in Sec.2, after a more precise problem definition (2.1), we describe the contour encoding algorithm which consists in a proper 1D signal sequence creation (2.2) followed by the wavelet coding of such a signal and the required side information (2.3). In Sec.3 we show some experimental results and for both shape coding (3.1) and generic contour map coding (3.2) Sec.4 provides some concluding remarks.

2 Encoding Algorithm

2.1 Problem Definition

We consider a *contour map* as a binary image where active pixels are in direct spatial relation with contour and/or shape information on the original image. A contour map can be defined as a set of non intersecting *contour tracts* or contour *segments*. A contour tract is a connected and non redundant (filiform), open or closed pixel sequence defined on the discrete spatial domain. Contour maps can be generated in various ways, for example by means of contour extraction operators, segmentation techniques, thinning or skeletonization algorithms.

In our framework the contour map can be interpreted as a single sequence of contour points and described by the corresponding sequence of cartesian coordinates. As already stated, this sequence is not at all unique because it depends on the scanning order of the single contour points and on the order the contour segments (if more than one) are concatenated. Supposing that the contour point scanning process on a single contour tract follows an adjacency criterion, the only signal discontinuities are generated by the coordinate jumps between subsequent contour segments (the coordinate differences between the end of a contour tract and the beginning of another one).

One can think wavelets are adequate to approximate discontinuities, however for accurate contour representation only moderate discontinuities can be tolerated in order to avoid the introduction of annoying artifacts. In fact, wavelet approximation smooths the signal discontinuity and can generate ringing artifacts near the discontinuity boundaries. The above effects cause false contour segment connections (due to signal smoothing) and a "crumbling" of the contour extremities (due to the "ringing") on the reconstructed contour map. These effects, other than be linked in their entity to the coding rate, are more or less visible depending on the value of the single coordinate jumps. The above observations are illustrated in Fig. 1 where a detail view of a low bit-rate (1 bit/contour point) sequence coding is shown: the larger (rightmost) discontinuity corresponds to strong oscillations in the reconstructed sequence, while the smaller (leftmost) discontinuity is completely smoothed and the corresponding contour tracts turn out to be joined. These impairing effects should be kept under control and therefore the discontinuity amplitudes on the 1D signal sequence must be controlled and possibly minimized. To do this we adopt specific solutions for handling and sequencing disconnected contour segments in order to improve both objective and visual coding performances.

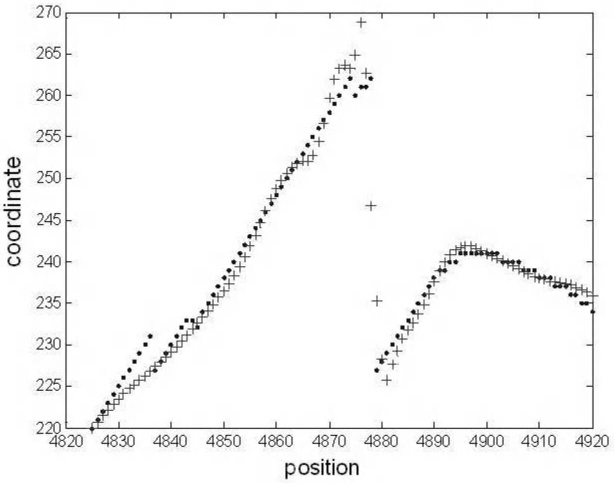


Fig. 1. Particular of a sequence of coordinates: with symbol "·" the original 1D signal sequence, with "+" a 1*bit/contour point* coded one

2.2 Creation of the One-Dimensional Sequence

Discontinuity handling has been achieved by the following two steps:

a. *global discontinuity minimization*: it consists in creating the 1D signal and finding an optimal or sub-optimal sequencing of individual contour segments,
b. *local discontinuity control*: it consists in treating the coordinate discontinuity between consecutive segments as side information when their amplitude exceeds a certain threshold.

The aim of the *global discontinuity minimization* (step a.) consists in minimizing the "total discontinuity path" which is covered by the coordinate signal. This mean finding a contour segment sequencing mechanism which minimizes the global amount of coordinate jumps. This problem can be set as a typical "Traveling Salesman Problem", with the inherent unmanageable complexity in finding the optimal solution. Sub-optimal suitable heuristics, which lead to the generation of contour segment linking paths, has been found and tested. Here we briefly describe the adopted solution which has been experimentally selected to be the most effective among other similar ones:

1. The first considered contour point is selected as the leftmost-upper one, and its position is stored.
2. according to a minimum path length criterion[1], the coordinates of the point that is nearest to the last added one are concatenated iteratively and generate two integer valued sequences $x(n)$ and $y(n)$ corresponding to the various abscissa and ordinates respectively; the total path T is computed as the sum of the euclidian distances (hereafter referred as "jumps") among adjacent contour points
3. The extremities $(n, n+1)$ of the greatest not yet considered jump are found, which correspond to two coordinates pair $(C(n), C(n+1))$, where a coordinate pair is defined by $C(n) = (x(n), y(n))$.
4. The right-hand sub-sequence $(C(n+1), ..., C(N_{right}))$ is found with $N_{right} > n+1$ such that $C(N_{right})$ is the contour point closest (in terms of euclidian distance) to $C(n)$; then the above sub-sequence is reversed (flipped) and the resulting new total path T_I is computed; in Fig. 2 this process is shown where $C(n) = A$, $C(n+1) = B$ and $C(N_{right}) = D$.
5. A dual processing with respect to 4. is performed for a left-hand sequence $(C(N_{left}), ..., C(n))$; in this case the total path T_{II} is computed.
6. $T_m = min(T, T_I, T_{II})$ is selected along with the corresponding coordinate sequences; one can observe that by maintaining the synchronization of the inversions on the abscissa and ordinate sequences no side information is needed when T_I or T_{II} are selected.
7. The process iterates on i from step 3., until, a stop criterion is achieved, for example, for a selected ϵ, $T_m(i-1) - T_m(i) < \epsilon$.

At this point (step b.), a *local control of the residual discontinuities* is needed. If a discontinuity is higher than a properly defined threshold it is convenient to remove it from the sequences of the coordinates. The positions of the discontinuities and the offsets are lossless entropy coded, memorized in a header of the bit-stream, and sent as side information to the decoder. In addition, the first point immediately subsequent to a discontinuity is removed from the sequence of contour point because it is redundant. The offset is then subtracted from the coordinates of all points which follow the same discontinuity. The signal sequences

[1] The nearest contour point which abscissa and ordinate are concatenated to the respective sequences is found by using a minimum 2D euclidian distance criterion. In case of two or more contour points at the same euclidian distance the contour point which better preserve the current contour direction is seleted.

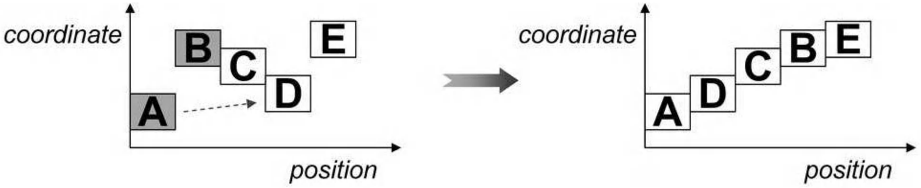


Fig. 2. Original sequence (on the left) and new generated sequence

of abscissas and of ordinates are also concatenated in order to obtain a single 1D signal. The threshold that is used for the local control of the residual discontinuities should be calculated to be optimal or near optimal in a rate-distortion (R-D) sense with respect to a proper distortion metrics. Because of its "structural" nature and of its direct influence on the visual results, an experimental threshold (e.g. found to be optimal with respect to a set of contour maps) could be used, in a reliable way, as an "invariant" of the method.

2.3 Wavelet Encoding with 1D-I-SPIHT

The obtained 1D signal and its side information perfectly represent the contour map. The 1D signal is ready to be encoded in the wavelet domain. Because of the peculiarity of the produced signals we experimentally tested various wavelet filter basis. The 16/4 biorthogonal spline basis demonstrated slightly better coding performance with respect to others and therefore it has been used for the results presented in Sec.3.

Thanks to the fact that wavelet transform does not modify the signal support, the proposed contour coding actually preserves the original number of contour points on all decoded contour maps. This turns to be a nice property of our scheme and facilitates the evaluation of the quality metrics considered hereafter.

To encode the wavelet coefficients we used a one-dimensional and improved version of the SPIHT [6] algorithm. We adopted the algorithm called I-SPIHT which has already been tested for 2D and 3D data [7]. In I-SPIHT some redundant bits (whose value can be deducted unambiguously) are removed and the arithmetic coding part has been improved. The interested reader can refer to [7] for a detailed explanation of the solutions adopted in I-SPIHT. The adaptation of the SPIHT or I-SPIHT algorithm to a one-dimensional transformed domain is straightforward and so details are omitted.

Heading and side information is entropy coded by using a simple Huffman encoder. The compressed header contains the total number of contour points, the 1D-I-SPIHT setup information and the x and y coordinates and amplitudes of the residual discontinuities. The size of the coded header on the total bitstream length depends on the amount and position of the contour tracts of the contour map to code. At 2 bit/contour point we experimentally determined a 1% ratio in case of shape coding while this ratio is about 20% in case of complex image contour map coding. The decoder first decode the header and then it progressively reconstruct an approximate version of the original shape or contour map.

3 Experimental Results

We tested the proposed technique on several contour-map data. Representative results are described here for shape coding and generic contour-map coding.

3.1 Shape Coding Results

For a performance evaluation on the shape coding case, we compare our results to those obtained with state-of-art techniques: the line-based methods "baseline-based" [2] and "vertex-based" [3] and the bitmap-based methods MMR (Modified Modified Read) [4] and CAE (Context-Based Arithmetic Encoding) [5]. In particular CAE is the solution adopted by the MPEG-4 standard. To evaluate the distortion on the reconstructed shape, we evaluate the D_n and D_p quality measure. These measures have been defined in MPEG-4 and used for shape coding performance evaluations [8]. D_n represents the number of erroneously represented pels of the coded shape divided by the total number of pels belonging to the original shape. D_p is the peak deviation, measured in pixel, where the deviation is calculated as the euclidian distance between the center of mass of a reconstructed pixel and the center of mass of the nearest original pixel. As test data we used the first frame of the test sequence "Kids", in SIF format (352x240 pels). Fig. 3 shows the total bits required for coding the shape as function of D_n. We also limited our analysis to the condition $D_p \leq 3$ which determines the right end point of the various curves of Fig. 3. In fact, it has been evaluated that a

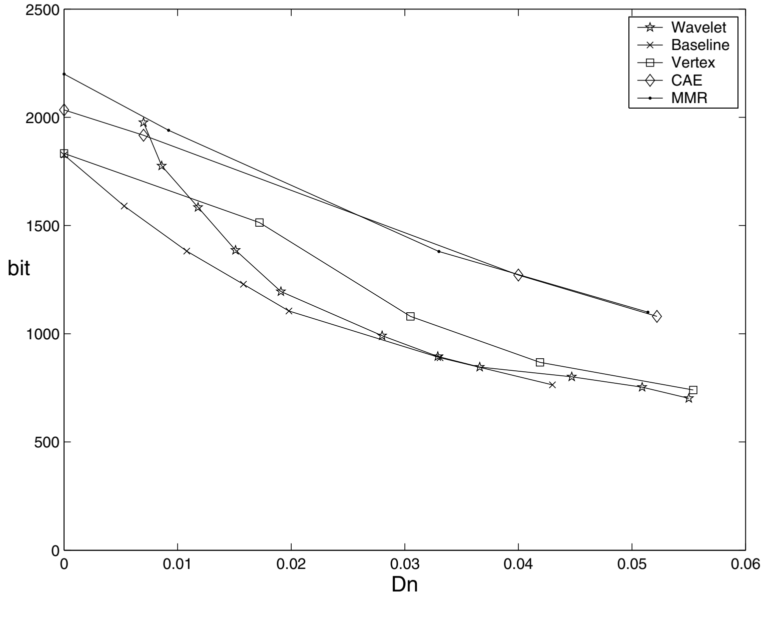


Fig. 3. Shape coding bits in function of D_n for the first frame of the sequence "Kids"; the proposed technique is labeled "Wavelet"

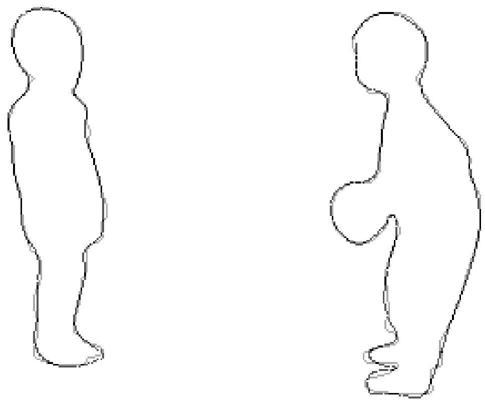


Fig. 4. Original shapes (black lines) and shapes encoded with 702 bits

$D_p > 3$ produces decoded shapes which are not suitable for video coding purpose [8]. For near-lossless rates the proposed technique loses efficiency. This problem is actually common to the considered contour-based techniques (and in general to most lossy contour coding algorithms). When a near lossless condition is required, the algorithms usually start to employ a chain-code. This solution could also be adopted for the described technique (or another solution explored). For lower bit-rates the proposed technique is more efficient than the 2 bitmap-based methods and also than the vertex-based technique; it approaches the performance of the baseline-based method that is currently the most efficient method reported in the literature. Moreover, the proposed technique reaches lower-rates

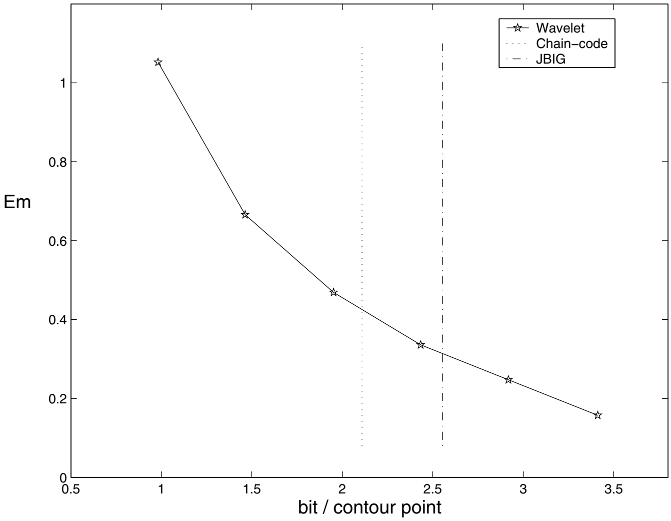
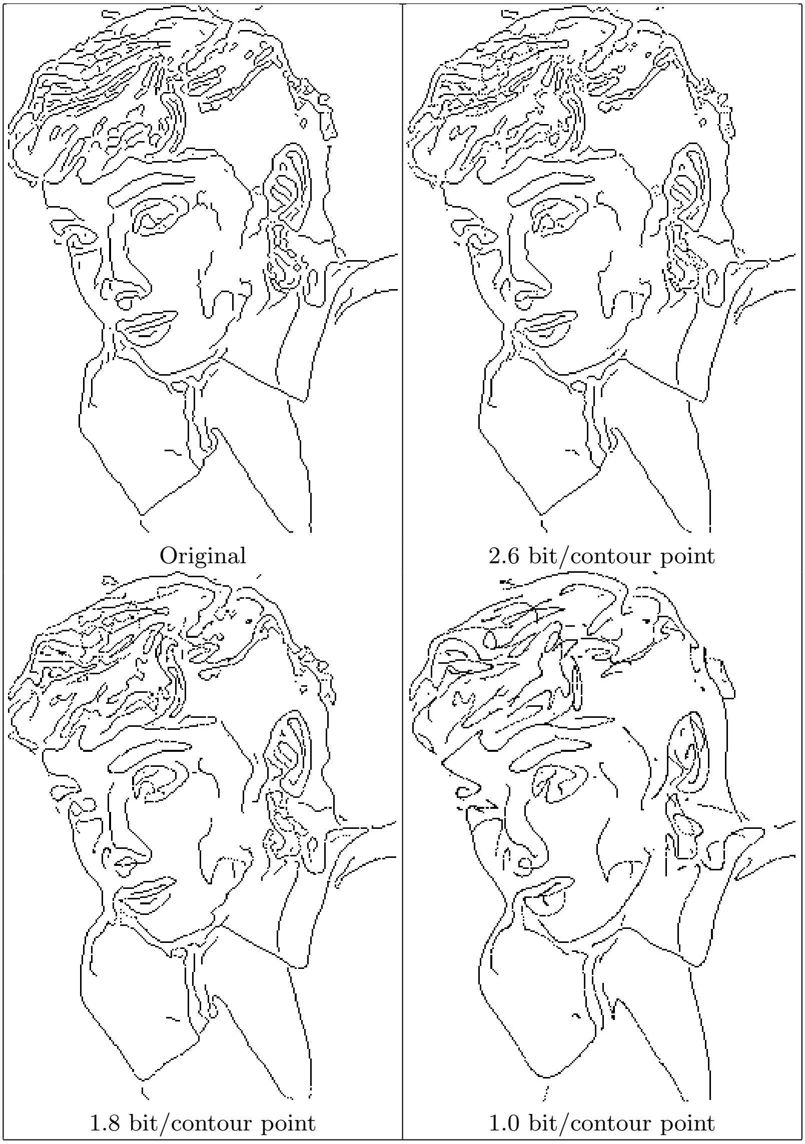


Fig. 5. Coding results in terms of Em for the "A.E." contour map. Lossless coding rates are indicated for the chain-code and JBIG techniques.

and higher D_n without violating the condition $D_p \leq 3$ pels. This can be interpreted as a good artifact control and actually corresponds to good perceptual quality performance at low bit-rates. Fig. 4 shows in grey line the decoded shapes at the minimal allowed bitrate (702 bits, $D_p = 3$) on top of the original ones (in black line). Another important aspect is that the proposed technique is the only one that produces a progressively decodable bit-stream.



Original

2.6 bit/contour point

1.8 bit/contour point

1.0 bit/contour point

Fig. 6. Visual coding results for "A.E." contour map

On this basis and other objective and visual tests we can conclude that, for the case of shape coding, our method presents performance comparable to state-of-art techniques and even at very high compression ratio (well under 1 bit/contour point) it doesn't introduce heavy or annoying artifacts while it preserves most details.

3.2 Generic Contour Map Coding Results

In this experiment we start from the contour map generated by the Canny method [9] on an image of "Audrey Hepburn". Coding results are shown in Fig. 5 in terms of E_m, i.e. the mean error (deviation) measured (in pixel/contour point). A *6 pixel distance* among the discontinuity coordinate extremities has been experimentally found a suitable threshold value for local discontinuity control (see Sec.2.2). In fact, it guarantees good visual quality performance and in most cases it minimizes (in a R-D sense) the mean error deviation E_m. Lossless rates are also reported as a reference for two lossless techniques: an 8-connectivity chain-code technique (line-based) [10] and the bitmap-based JBIG standard (bitmap-based) [11]. In this case multiple open contours must be represented and this generates a certain amount of side information. Visual results are very encouraging as shown in Fig. 6. Even at bit-rates of about 1 bit/contour point no annoying artifacts have been observed.

4 Conclusions

In this paper we proposed a new progressive shape coding technique based on the wavelet transform, suitable for any kind of contour images. The use of the wavelet transform, the unconstrained contour coding capabilities, the progressive structure of the coded bit-stream are the most original aspects of the proposed method, while the obtained coding performance makes this technique suitable for different image and video analysis and representation applications. The possibility of progressively decoding simple shapes or entire contour-maps can be useful or even required in many concrete situations (e.g. preview generation), it enables a fine-grain quality scalability and makes it easy to implement an unequal error protection. Progressive decoding is not possible in all the other considered techniques. In the case of shape coding, simulation results show that the proposed technique is as efficient as the best techniques reported in literature, while for application scenarios of lossy contour map coding perceptual results are already encouraging. Moreover, further optimizations of the entropy coding part are possible while other useful bit-stream properties, such as spatial scalability, seem to be achievable and are under consideration for future work.

References

1. Zahn, C. T., Roskies, R. Z.: Fourier Descriptors for Plane Closed Curves. IEEE Trans. Computers 21(**3**) (1972) 269–281
2. Lee, S., Cho, D., Cho, Y., Son, S., Jang, E., Shin, J., Seo,Y.: Binary Shape Coding Using Baseline-Based Method. IEEE Trans. Circ. and Sys for Video Technol. 9(**1**) (1999) 44–58

3. O'Connell, K.J.: Object-adaptive vertex-based shape coding method. IEEE Trans. Circ. and Sys for Video Technol. 7(**1**) (1997) 251–255
4. Yamaguchi, N., Ida, T., Watanabe, T.: A binary shape coding method using modified MMR. In: Proc. of Int. Conf. on Image Proc. (1997) 504–508
5. Brady, N., Bossen, F., Murphy, N.: Context-based arithmetic encoding of 2D shape sequences. In: Proc. of Int. Conf. on Image Proc. (1997) 29–32
6. Said, A., Pearlman, W. A.: A new, fast, and efficient image codec based on Set Partitioning in Hierarchical Trees. IEEE Trans. Circ. and Sys for Video Technol. 6(**3**) (1996) 243–250
7. Signoroni, A., Arrigoni, M., Lazzaroni, F., Leonardi, R.: Improving SPIHT-based compression of volumetric medical data. In: Proc. of Picture Coding Symp. (2001) 187–190
8. Katsaggelos, A. K., Kondi, L. P., Meier, F. W., Ostermann, J., Schuster, G. M.: MPEG-4 and rate-distortion-based shape-coding techniques. Proceedings of the IEEE 86(**6**) (1998) 1126–1154
9. Canny, J.: A computational approach to edge detection. IEEE Trans. on Pattern Anal. and Machine Intell. 8(**6**) (1986) 679–698
10. Eden, M., Kocher, M.: On the performance of a contour coding algorithm in the context of image coding - Part I: Contour segment coding. Signal Processing **8** (1985) 381–386
11. Kuhn, M.: JBIG-KIT lossless image compression library. http://www.cl.cam.ac.uk/∼mgk25/jbigkit/

Description of Evolutional Changes in Image Time Sequences Using MPEG-7 Visual Descriptors

Lei Ye[1], Lingzhi Cao[2], Philip Ogunbona[1], and Wanqing Li[1]

[1] University of Wollongong, Australia
{lei, philipo, wanqing}@uow.edu.au
[2] Zhengzhou University of Light Industry, China
caolingzhi@zzuli.edu.cn

Abstract. Colour and texture visual descriptors have been developed to represent structural features of images, mainly under the Query-by-Example (QBE) image retrieval paradigm. This paper explores applicability of MPEG-7 visual descriptors to describe and measure evolutional changes in image time sequences, using a fruit rotting process as an example. The research found that MPEG-7 visual descriptors can be applied to describe evolutional changes in image time sequences. The experimental results are provided using bananas captured in image time sequences. The results show the desirable monotonicity of description metrics of MPEG-7 similarity matching for image time sequences and their sensitivity to practical image acquisition conditions. Our experiments demonstrate that Colour Layout descriptors (CLD) and Scalable Colour descriptor (SCD) describe the changes that are consistent to the degree of visual changes while CLD and Homogeneous Texture descriptor (HTD) are more robust to variations of image data due to practical image acquisition conditions.

1 Introduction

With the proliferation of multimedia content and advances in multimedia storage and communication technologies, the management of the multimedia content becomes the focus of recent research. ISO/IEC MPEG (Moving Picture Expert Group) committee has standardised a set of visual descriptors based on structural information of the images, known as Multimedia Description Interface or MPEG-7 [1]. Each visual descriptor describes certain aspects of images. For instance, colour descriptors describe dominant colour, colour layout and colour structures of images; texture descriptors describe image regularity; shape descriptors describe region-based and contour shapes in images. There are some other descriptors to describe videos. The standardized MPEG-7 visual descriptors [2] have undergone vigorous testing for their expressiveness under the visual similarity Query by Example (QBE) paradigm [3][4]. It is advantageous to use standardised visual descriptors for various applications, as they will be exchangeable and available in MPEG-7 aware multimedia systems.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 189–197, 2006.

© Springer-Verlag Berlin Heidelberg 2006

As a standard description tool, visual description can be explored for a wide range of applications besides QBE in image retrieval systems. In this paper, various visual descriptors are investigated for description of object status in image time sequences. Some objects of interests in the image may change over time. For example, fruits may rot and wine may mature. In these evolutional processes, changes of object status are often observed by their colours, textures or distributions of them. For each descriptor, a metric is recommended by the MPEG-7 to measure the distance between visual descriptions of two images [5]. These metrics are initially proposed for image retrieval.

In this paper, various visual colour and texture descriptors and their metrics are explored for describing the evolutional change of the object and measurement of such changes. In many applications of this kind, monotonic increasing of the description metrics over the image time sequences is a desirable property. We believe this is the first time that MPEG-7 visual tools is used to describe object status of an evolutional process captured in image time sequences.

Section 2 of the paper describes colour and texture visual descriptors and their similarity metrics. Section 3 presents the experimental results on banana image time sequences.

2 Colour and Texture Visual Descriptors

MPEG-7 visual description tools consists of basic structures and descriptors that cover basic visual features. Four visual descriptors are used in this research. A full description of these descriptors, their extraction processes and similarity metrics are given in [2][5][6].

2.1 Dominant Colour Descriptor (DCD)

This descriptor specifies a set of dominant colours either for the whole image or for any arbitrary shaped region. The dominant colour extraction algorithm takes as input a set of pixel colour values specified in the RGB colour space. A spatial coherency (SC) on the entire descriptor is also defined that specifies the spatial coherency of the dominant colours described by the descriptor. It is computed as a single value by the weighted sum of per-dominant-colour spatial coherencies.

The similarity metric for the descriptor is described as follows.

$$D(DCD_1, DCD_2) = w_1 \cdot |SC_1 - SC_2| \cdot DC\_Diff + w_2 \cdot DC\_Diff \ , \qquad (1)$$

where SC_1 and SC_2 are spatial coherencies normalized from 0 to 1. DC_Diff is the difference between two sets of dominant colours, w_1 is the weight of the first term and w_2 is the weight of the second term.

2.2 Scalable Colour Descriptor (SCD)

This descriptor is a colour histogram in HSV colour space, which is encoded by a Haar transform. The feature extraction consists of a histogram extraction in HSV colour space, uniformly quantised into 256 bins according to the tables provided in the normative parts, and histogram values then nonlinearly quantised.

The l_1 norm in either the coefficient domain or the histogram domain is used as the similarity metric for this descriptor.

2.3 Colour Layout Descriptor (CLD)

This descriptor represents a spatial distribution of colours. The descriptor is extracted from the 8 × 8 array of local representative colours of the image with DCT transformation applied to three colour components (Y, Cb and Cr). CLD is resolution-invariant.

The similarity metric between two descriptor values $CLD_1(Y_1, Cb_1, Cr_1)$ and $CLD_2(Y_2, Cb_2, Cr_2)$ is calculated as follows.

$$\begin{aligned} D(CLD_1, CLD_2) = &\sqrt{\sum_i w_{yi}(Y_{1i} - Y_{2i})^2} \\ &+ \sqrt{\sum_i w_{bi}(Cb_{1i} - Cb_{2i})^2} \\ &+ \sqrt{\sum_i w_{ri}(Cr_{1i} - Cr_{2i})^2} \ , \end{aligned} \tag{2}$$

where the subscript i stands for the zigzag scanning order of the coefficients.

2.4 Homogeneous Texture Descriptor (HTD)

Texture represents the regularity of an image such as directionality, coarseness, regularity of patterns etc. This descriptor characterizes the region texture using energy and energy deviation values from a frequency layout that constitute a homogeneous texture vector. The first and the second components of the descriptor are extracted by computing the mean and standard deviation of the image pixel intensities. The remaining energy and energy deviation features are computed by applying a set of 30 Gabor filters (6 orientations and 5 scales) in the frequency domain.

The similarity metric for this descriptor is described as follows.

$$d(HTD_1, HTD_2) = \sum_k \left| \frac{TD_1(k) - TD_2(k)}{\alpha(k)} \right| \ . \tag{3}$$

The recommended normalization value $\alpha(k)$ is the standard deviation of $TD(k)$ for a given database, in this case the image set in the sequence or user's own choice.

3 Description of Object Status in Image Time Sequences

Some research has been reported to use image processing [7][8], artificial neural network [9] and machine vision [10] to assess the status of fruits in various applications such as intelligent fridges, fruits condition monitoring system and automatic fruit grading systems.

In our common knowledge, the rotting process of bananas can be visually observed by their changes of colour and textures. Therefore, colour and texture descriptors are used to measure the visual changes of the banana image time sequences. Colour is an important indicator of the banana status. As the bananas rotten, the colours change. In fact, new colours emerge and the layout or structure of the colours changes. As a result, the texture of the object changes as well.

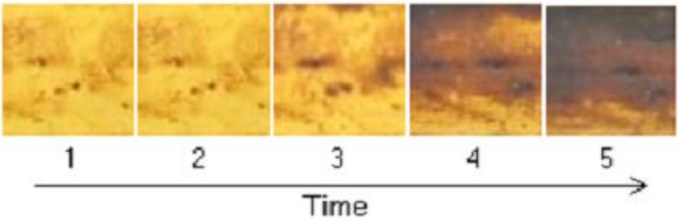


Fig. 1. Visual Changes of a Banana Local Area over Time

First experiment investigates the applicability of MPEG-7 colour and texture descriptors. Images taken at 5 different time points of a local area of a banana are used. This investigation compares the distance measurement of various visual descriptors and visual changes observed by human eyes. Figure 1 shows a local area of a banana at different time points. Figure 2 shows the normalised distances of these changes measured by MPEG-7 visual descriptors with the first image as the reference.

Visually, the change of the banana from the time point 1 to the time point 2 is insignificant. At the time point 3, the banana starts rotting and the visual change is significant comparing with that at the time points 1 and 2. The banana at time points 4 and 5 is rotten and the visual change is very large. From the Figure 2, CLD and SCD shows small distances between time points 1 and 2 while HTD and DCD shows large distances between them, which is sensitive to minor visual changes. At the time point 3, CLD and SCD shows a steady increase comparing to the time points 1 and 2 while HTD and DCD show smaller increases comparing to the their increases at the time point 2. At the time point 4 when the banana is rotten, all of them show large distance increases while the increases of CLD and SCD are more dramatic. It appears that CLD and SCD are more consistent to visual changes.

The next experiment applies visual descriptors to two sets of image sequences of bananas consisting of 10 sequential images. The two image sequences of bananas used in the experiment are shown in Appendix. There are minor noticeable variations of images in the same sequence. The focus of the camera is not exactly the same due to the camera's auto focus function that results in slight differences in individual shots. There are noticeable displacements of the banana

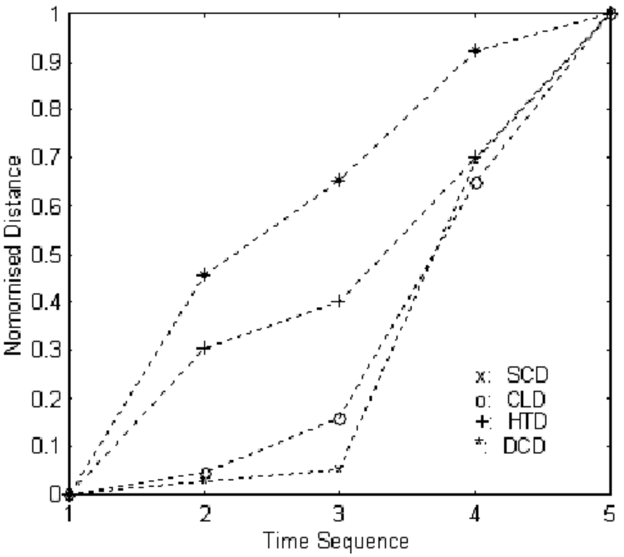


Fig. 2. Normalised Distances of Visual Changes Measured by Visual Descriptors

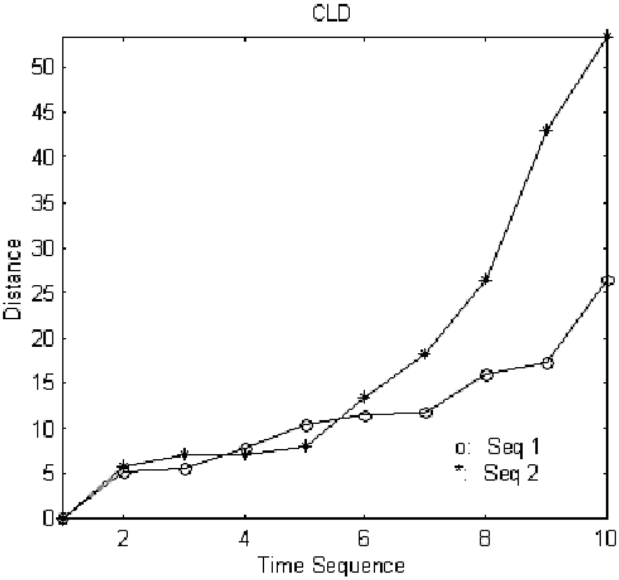


Fig. 3. CLD Distances of Image Sequences

in the images, that is, they are not pixel-by-pixel aligned from one image to another. It is expected that there are minor luminance differences as they are taken at different time over several days. However, this is a realistic condition in practice.

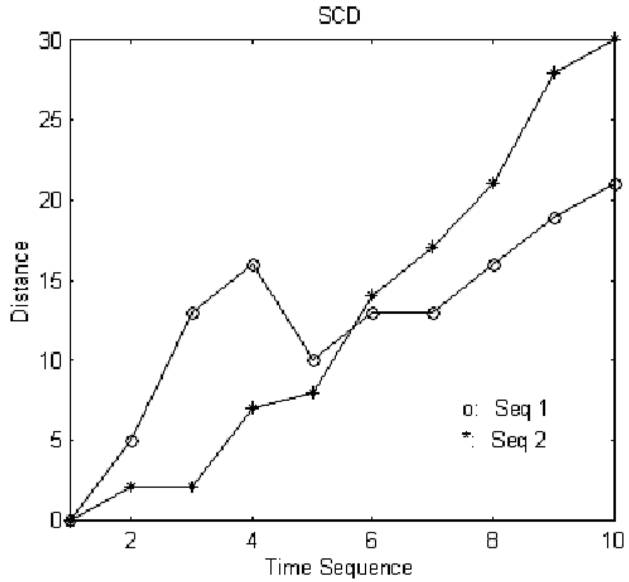


Fig. 4. SCD Distances of Image Sequences

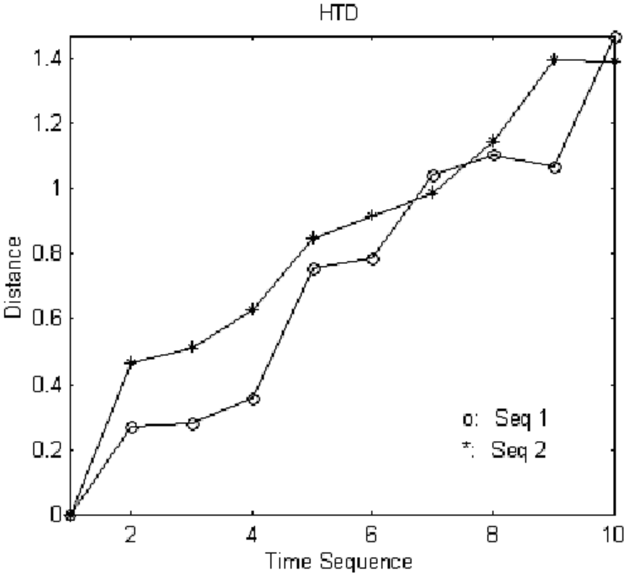


Fig. 5. HTD Distances of Image Sequences

Figures 3 to 6 show the distances of the images in the sequences with the first image as the reference using CLD, HTD, DCD and SCD, respectively. CLD shows the most consistent monotonicity of the metric for both image sequences. HTD shows consistent monotonicity of the metric for Sequence 2 but slight dip

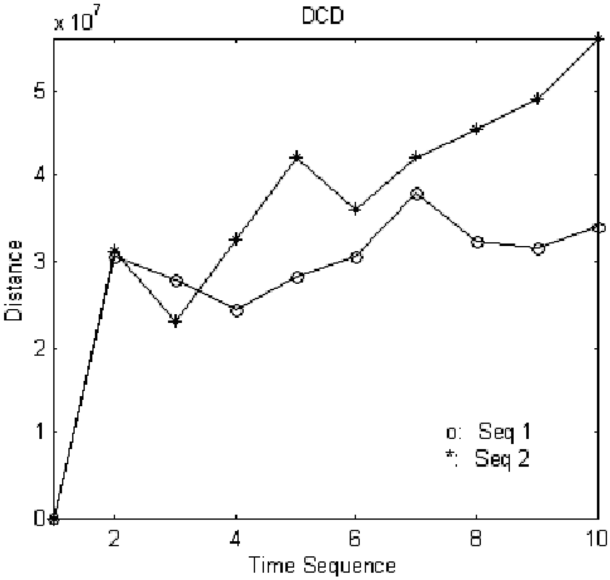


Fig. 6. DCD Distances of Image Sequences

at the time point 9 for sequence 1. Visual inspection found noticeable variations in the image at the time point 9, which has no significant impact on CLD. SCD shows consistent monotonicity of the metric for Sequence 2 but inconsistent for Sequence 1. DCD shows no consistent monotonicity of the metric for both sequences.

In summary, both CLD and HTD metrics show a steady trend consistent to the banana rotting process over the time and CLD is the most robust to the variations of conditions.

4 Remarks and Future Work

The standard visual descriptors are primarily developed to represent structural information of images for image retrieval under QBE paradigm. The various visual descriptors are supposed to describe certain aspects of the visual features of images. This research investigated their applicability to describe evolutional process of objects in the image time sequences. It is found that all four visual descriptors are able to describe the visual changes with monotonic increasing of the distances recommended by MPEG-7 over the time for well-aligned image sequences. It is also found that the distance metrics of CLD and SCD are more consistent to the degree of the perceived visual changes as shown in Figure 2. However SCD and DCD are more sensitive to the variations of focus and displacements resulted in the practical acquisition process of image time sequence. CLD and HTD demonstrate more robustness for those variations, as shown in Figures 3 to 6.

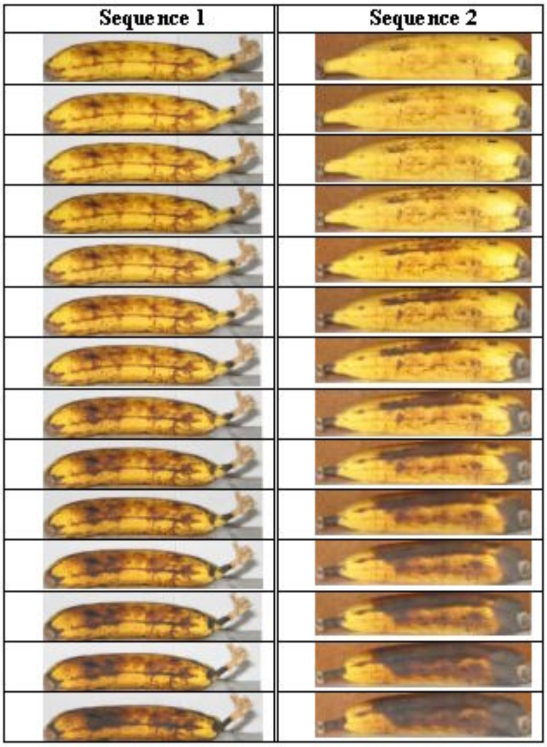
Some criteria can be developed to determine the stages or the turning points of the evolutional processes of fruits. Our experience shows that differential or aggregated distances can be used for this purpose though they are application dependent.

The detailed investigation of various MPEG-7 visual descriptors in respect to their sensitivity to different variations of the image quality would be an interesting future work. As the MPEG-7 similarity metric, which is an informative part of the standard, is meant for image retrieval, new metrics for visual descriptors for specific applications can be developed, in particular, metrics that can lead to determine the stages of the object evolutional processes.

References

1. Martinez, J., MPEG-7 Overview. ISO/IEC JTC1/SC29/WG11, N5525 (2003).
2. ISO/IEC 15938-3:2001. Multimedia Content Description Interface - Part 3: Visual.
3. Ndjiki-Nya, J., et al.: Subjective Evaluation of the MPEG-7 Retrieval Accuracy Measure (ANMRR). ISO/IEC JTC1/SC29/WG11, M2029 (2000).
4. Manjunath, B. S., et al.: MPEG-7 color and texture descriptors. IEEE Trans. Circuits Syst. Video Technol. **11** (2001) 703–715.
5. ISO/IEC 15938-8:2001. Multimedia Content Description Interface - Part 8: Extraction and Use of MPEG-7 Descriptions.
6. Manjunath, B.S., et al.: Introduction to MPEG-7. Wiley (2002).
7. Njoroge, J.B., et al.: Automatic Fruit Grading System using Image Processing. Proc. Of the 41st SICE Annual Conf **2** (2002).
8. Chan, W. H., et al.: Vision based fruit sorting system using measures of fuzziness and degree of matching. IEEE International Conference on Systems, Man, and Cybernetics **3** (1994).
9. Morimoto, T., et al.: Optimization of storage system of fruits using neural networks and genetic algorithms. Proceedings of 1995 IEEE International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium **1** (1995) 20-24.
10. Aleixos, N., et al.: Assessment of citrus fruit quality using a real-time machine vision system. Proceedings. 15th International Conference on Pattern Recognition **1**(2000).

Appendix: Image Time Sequences



Per-GOP Bitrate Adaptation for H.264 Compressed Video Sequences

Fabio De Vito[1,⋆], Tanir Ozcelebi[2,⋆⋆], Reha Civanlar[3], A. Murat Tekalp[4], and Juan Carlos De Martin[5]

[1] Dip. Automatica e Informatica, Politecnico di Torino, Torino 10129, Italy
Dep. of Electrical Engineering, Koc University, Sariyer, Istanbul 34450, Turkey
fabio.devito@polito.it, fdevito@ku.edu.tr
http://media.polito.it/devito

[2] Dep. of Electrical Engineering, Koc University, Sariyer, Istanbul 34450, Turkey
tozcelebi@ku.edu.tr
http://home.ku.edu.tr/∼tozcelebi

[3] Dep. of Electrical Engineering, Koc University, Sariyer, Istanbul 34450, Turkey
rcivanlar@ku.edu.tr

[4] Dep of Electrical Engineering, Koc University, Sariyer, Istanbul 34450, Turkey
Dep. of El. and Comp. Eng., Un. of Rochester, Rochester, NY 14627-0126, USA
mtekalp@ku.edu.tr

[5] Dip. Automatica e Informatica, Politecnico di Torino, Torino 10129, Italy
demartin@polito.it
http://media.polito.it/demartin

Abstract. In video transmission over packet data networks, it may be desirable to adapt the coding rate according to bandwidth availability. Classical approaches to rate adaptation are bitstream switching, requiring the storage of several pre-coded versions of a video, or layered (scalable) video coding, which has coding efficiency and/or complexity penalties. In this paper we propose a new GOP-level rate adaptation scheme for a single stream, variable target bitrate H.264 encoder; this allows each group of pictures (GOP) to be encoded at a specified bitrate. We first compare the performance of the standard H.264 rate control algorithm with the proposed one in the case of constant target bitrate. Then, we present results on how close the new technique can track a specified per-GOP target bitrate schedule. Results show that the proposed approach can obtain the desired target rates with less than 5% error.

1 Introduction

In recent years, technological improvements have created the basis for the development of several new network applications, among which the most challenging one seems to be the access to media contents. In the case of multimedia transmission across packet-switched data networks, channel throughput variations can

⋆ Supported by Centro Supercalcolo Piemonte (CSP) - Torino, Italy.

⋆⋆ Supported by 3DTV-NoE funded by the European Commission 6^{th} Framework Information Society Technology Programme.

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 198–206, 2006.

© Springer-Verlag Berlin Heidelberg 2006

affect the quality of the received information. This is especially true in the case of wireless communications, where the channel varies continuously according to the number of users and their relative position and speed with respect to the base station. If insufficient bandwidth is available, the playout quality may be lowered by the presence of packet losses; in this case, it is preferable to recode the stream at a bitrate lower than the available throughput, rather than relying on correction capabilities of concealment algorithms. On the other hand, if more bandwidth is available, the media content can be safely coded at higher rate and experience no losses, so increasing the quality.

Network conditions are not the only factors which can impose a modification in the target rate: if the user accesses the contents paying on a per-byte basis, he can desire to receive a low-quality stream when low-importance contents are played, and require better coding when the stream contains sequences he considers as high-importance.

In case of video transmission over packet data networks, modern video codecs like H.264 [1] can achieve very low bitrate coding of sequences. The use of such coders allows the distribution of video contents also on low-bandwidth links, at rates which are usual in wireless communications.

Unfortunately, the radio link suffers of wide bandwidth oscillations and, in particular at very low bitrates, concealment algorithms do not guarantee a satisfactory recovery of the eventually lost information, so quickly degrading the perceived video quality; moreover, when high compression is required, packets usually contain an entire frame and a loss has impact on a large number of macroblocks (MB). Recoding at lower rate in this case would introduce intentionally a (controllable) error on nearly all pixels; instead, if losses occur, the residual concealment distortion will be on average higher than in the recoding case, and moreover it will be concentrated in the frames interested by the loss, and sometimes uncontrollably.

For the above reasons, adaptivity of video streams has been extensively studied in recent years. Mainly, the adaptation is achieved by storing several versions of the same content, encoded at different bitrates, and then switching among the streams as required by the network condition. This approach is particularly suitable for video archives, where the access is *on demand* and there is enough time to perform multiple encodings. The server can then transmit at the appropriate bitrate according to the network available bandwidth. However, resynchronization issues may arise when switching.

Another well-known approach to the same problem is layered video coding (see, e.g., [2]). This consists mainly in coding a *base layer* at low quality, and then adding one or several *enhancement layers*. The receivers are able to decode the base layer independently, and the reception of one or more enhancement layers can refine the video quality: the bitrate adaptivity is obtained by changing the number of enhancement layers transmitted [3]. An extreme case of this technique is *fine grain scalability* [4, 5], which allows a very small error in bitrate adaptation.

Both the above approaches demonstrated to be effective in achieving good network utilization and high video quality [6, 7]. However, some of them can only achieve bitrates in a limited set, usually decided at coding time; for example, it is limited to the rates of the pre-coded versions in case of simultaneous storage, while it is given by the number of enhancement layers transmitted in case of layered video coding.

In literature it is possible to find several papers overviewing the above concepts (e.g.[8]), and extending them with techniques like frame skipping or coefficient dropping [9, 10, 11].

In this work we propose modifications to the standard H.264 bitrate control routine, in order to make the stream change its bitrate to an arbitrarily chosen value *on the fly*. With respect to the above mentioned approaches, this solution will output one single, non-layered stream; its GOPs can be coded each one at a different rate, achieved with high precision, therefore increasing the granularity and avoiding simultaneous storage of pre-coded information.

The paper is organized as follows. In Section 2 we describe the background to the problem, and in Section 3 we outline the modifications implemented within the encoder. We compare coding results in Section 4, and draw the conclusions in Section 5.

2 Background

In this work, we make reference to functionalities implemented within the JM 9.3 H.264 standard codec. This modern video codec is able to achieve very low bitrate values, due to its advanced redundancy reduction capabilities. The reference software implements a rate control algorithm, which requires as input the target value and a starting quantization parameter for the first I-frame of the sequence. The output will be a constant bitrate (CBR) sequence, usually it will converge to the desired value after a period of one or two GOPs and it will fluctuate around the target with reasonable approximation for the remaining part of the sequence. This standard rate control is useful to create a single sequence at a given bitrate. It is not possible to change the bitrate during the coding operation, and even if it was possible, the convergence time of a couple of GOPs would allow only a very low frequency in changes to get meaningful results.

Our goal will be modifying this rate control system to produce a single stream, encoded according to a per-GOP bitrate pattern; the values of this pattern may be produced either by some bandwidth estimation tool, to adapt to changed link conditions, or by the user who desires lower or higher quality according to his own preferences and needs, or both of them. In theory, this change in the desired bitrate can occur at *each* single GOP boundary, and so the convergence speed will become a key issue, in order to reach the per-GOP target value before the following switch is required.

Due to the main bitrate algorithm, changes in the target can occur only when an I-frame is reached. As it is implemented in the reference codec, the GOP length is fixed by indicating the periodicity of I-frames and the number of

B-frames in a run. If the number of frames within a GOP is small and the frame rate high enough, this fixed behavior does not represent an obstacle for adaptation to the network conditions, since there can be several GOPs starting in a second of video and the granularity of the switch points for most of the applications can be considered satisfactory. Typically, at very low bit rates, the frame rate is usually reduced to values around 15 fps or lower, and the GOP length is usually high in order to mitigate the presence of I-frames, which require more bits to be coded with respect to P-frames. This forces switches to occur every two or three seconds, which may be not useful to compensate a fastly changing network throughput. If we desire to switch with high frequency, then also the selection of I-frame position should be changed. The constraint on fixed GOP structure should be relaxed to gain more flexibility in this case. It is possible to modify the length and structure of each GOP dynamically without affecting the decodability of the sequence, since the decoder is able to operate with any I-/P-/B-frame pattern, regardless of the structure of previous GOPs.

If the two above described functionalities are active jointly, it will be possible to remotely drive the encoder via a simple socket program, by communicating at which frame the switching should occur and to what value of bitrate.

In this work, we will focus our attention only to the bitrate control algorithm, and we do not show the effect of variable GOP sizes. Thus, we will refer to 1-second GOPs, containing 30 frames when working at 30 fps and 15 frames when working at 15 fps, if not differently specified.

3 GOP-Level Rate Adaptation Scheme

Being the desired effect to switch between bitrates within the same sequences, more flexibility in the definition of this parameter is required: in our implementation, the codec may read from a socket or a file the new value of bitrate at each beginning of a GOP (I-frame). This approach is suitable for a remotely-driving of the codec by simply employing a socket communication system. In the same way, the encoder can receive the desired length of the GOP being coded, so tuning also the position of I-frames. Once the new desired bitrate is read and stored in memory, the standard rate-control routine will automatically converge to this new value.

It is still necessary to speed up the convergence to the specified bitrate. It is possible to have one bitrate switching request for each GOP and the standard encoder could not be able to converge to the new value in such a short time. Furthermore, the codec stores internally some statistics on previous GOPs, which become useless, and meaningless, when the target rate is modified; those statistics need to be tuned accordingly.

To ensure better convergence, we propose to recompute the initial quantization parameter for each GOP as described below. We implemented a static initial table, showing an approximate mapping between the quantization parameters (QP) and the bits per pixel obtained for that QP value. Every time a new I-frame is being coded, the desired bitrate is read and the target bits-per-pixel (*bpp*) indicator is computed according to the frame size and frame rate:

$$bpp(i) = br(i) \times \frac{len_{GOP}(i)}{fps} \times \frac{1}{w \cdot h \cdot 1.5}, \tag{1}$$

where $br(i)$ is the bitrate of the i-th GOP, len_{GOP} is the number of frames within the current GOP, fps is the frame rate (frames per second) and w and h are respectively frame width and height. The factor 1.5 takes into account the presence of the two sub-sampled chrominance components.

The initial quantization parameter is then chosen from the table as the one ensuring the closer bpp indicator. Every time a GOP terminates, and right before starting the following I-frame, the bpp obtained for the last GOP is stored in the table together with its average quantization parameter, so updating the starting *static* values at each step to better fit over the sequence characteristics.

This approach will produce better convergence also for the first GOP in the sequence, with respect to the standard implementation, because the initial quantization parameter is no longer required as input but internally determined; moreover, it will continue producing better results during encoding due to the dynamic update. The initial table does not need to provide the exact matching between bpp and quantization parameter QP because if the GOP contains a sufficient number of frames, the rate control algorithm will converge in any case to the desired target, after some inter-GOP oscillations.

The bpp values chosen for the initial table are shown in Table 1, where we present the portion of the table for $QP \in [25 - 50]$. We include values which are not low-bitrate to show that this approach works for a wide range of values.

The values shown in Table 1 have been obtained by setting $bpp = 4.73$ for $QP = 0$; this value has been chosen after a study of coding statistics for different sequences. All the other values in the initial table are obtained recursively by using (2):

$$bpp_{i+1} = bpp_i \cdot 0.9 \tag{2}$$

In Table 1, bits per pixel are intentionally computed ignoring the presence of chrominance components. This will cause the algorithm to choose a smaller starting quantization parameter, so letting a better quality for the I-frame. The bitrate convergence routine will then take care of quantizing more the following

Table 1. Initial table, some quantization parameters *QP* and bits per pixel *bpp*; for reference, the bitrate obtained with a QCIF frame size at 25 fps is also shown

QP	bpp	rate [bps]
25	0.368182	233280
30	0.220909	139967
35	0.132545	83980
40	0.079527	50388
45	0.047716	30232
50	0.028630	18139

Table 2. Comparison between bitrate achievement of standard H.264 JM 9.3 codec (Std) and the modified (Mod) version; the first GOP is excluded

Sequence	Error (%)					
	32 kbps		64 kbps		128 kbps	
	Std.	Mod.	Std.	Mod.	Std.	Mod.
Foreman	1.97	0.94	0.54	0.66	0.54	0.55
Tempete	0.70	0.85	0.22	0.41	0.43	0.25
Paris	2.61	1.20	1.84	0.87	1.16	0.29
News	1.44	1.02	0.64	0.26	0.33	0.38
Mobile	2.85	1.20	1.28	0.62	0.54	0.32

frames to match the bitrate. Good results have been obtained even with small number of frames per GOP.

To show convergence accuracy, we report the average error obtained by the standard H.264 JM 9.3 encoder and the modified version in Table 2, for a constant bitrate encoding of the sequences, at three different values.

This table shows the average percent error for different sequences encoded at different bitrates, using both the standard and the modified encoder. GOPs contain 30 frames each. The first GOP is excluded from the computation because the error obtained with the standard encoder is excessively high, even if, to speed up convergence, the quantization parameters for the first frame in the case of standard encoder have been set to 30, 37 and 45 respectively for the cases of 32, 64 and 128 kbps, which are values very close to the ones reported in the initial table of the modified coder (see Table 1 for comparison). Results show that the new encoder can achieve at least the same precision of the standard one, outperforming it for the majority of sequences and bitrates considered. Furthermore, we obtain the desired bitrate avoiding the two-GOP convergence time.

4 Results

The described modified H.264 encoder has been employed to code different video sequences over different bitrate patterns, which have been chosen, as limit case, to force a switching at every GOP boundary. In this section we present coding results for two sequences at 30 frames per second and for four sequences at 15 fps.

Fig. 1 shows the required bitrate pattern and the output that is obtained for the two sequences *foreman* and *mobile*. This pattern contains target values in the set $\{32, 64, 128\}$kbps. This is again a limit setting, since usually sequences at low bitrate are coded using sensibly less than 30 frames per second to gain better PSNR.

The plots result close to the reference; we show the percent error in Table 3 for each one of the nine coded GOPs of the two sequences. This error results to be never higher than 5%; as a consequence of imposing several changes within the sequence, these values are higher than the ones shown for the modified coder in Table 2.

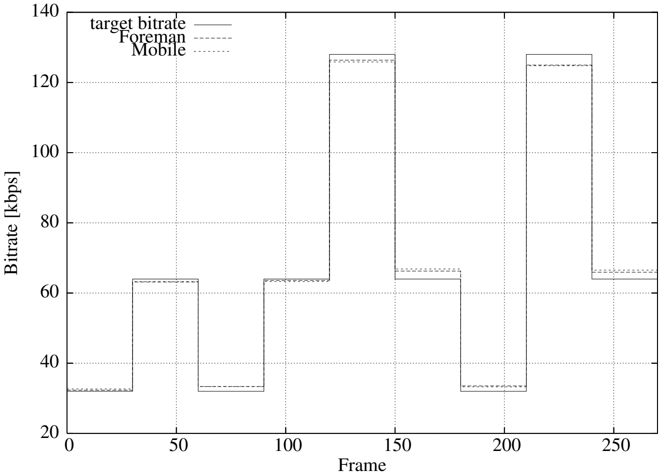


Fig. 1. Bitrate obtained for two sequences at 30 frames per second, compared with the target bitrate behavior

Table 3. Error in achieving target bitrates as shown in Fig. 1, for two sequences at 30 frames per second

GOP	Error (%)	
	Foreman	Mobile
1	0.97	2.05
2	1.34	1.21
3	4.28	4.10
4	0.66	1.14
5	1.28	1.69
6	3.50	4.45
7	4.78	3.90
8	2.38	2.51
9	3.00	3.94

Better results can be obtained coding sequences at lower bitrates and at lower frames per second. We encoded four sequences at 15 fps, requiring a very low bitrate value. In this scenario we set the required rate to switch between 25 and 35 kbps. The resulting behavior for each sequence is reported in Fig. 2. Again, the performance of the rate control routine follows the target pattern. The percent errors placed in Table 4 are smaller than the ones reported in Table 3 and comparable with the ones of Table 2, due to the smaller difference in the values among which we switch. Sequence *Tempete* is shorter and contains only 8 GOPs.

These results show that the proposed implementation can achieve the effect of encoding portions of a single sequence following a specified behavior from close up.

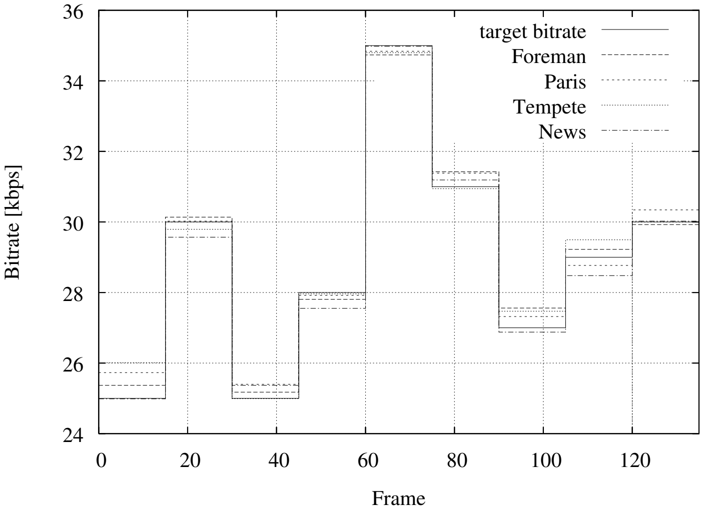


Fig. 2. Bitrate obtained for four sequences at 15 frames per second, compared with the target bitrate behavior

Table 4. Error in achieving target bitrates as shown in Fig. 2 for four sequences at 15 frames per second

GOP	Error (%)			
	Foreman	Tempete	Paris	News
1	1.47	4.03	2.91	0.06
2	0.45	0.69	0.08	1.44
3	0.70	0.03	1.60	1.47
4	0.69	0.14	0.29	1.60
5	0.75	0.57	0.46	0.05
6	1.37	0.18	1.24	0.62
7	2.07	1.75	1.19	0.44
8	0.77	1.71	0.80	1.79
9	0.24		1.15	0.08

5 Conclusions

In this paper, we proposed some simple modifications to the standard H.264 JM codec, version 9.3, in order to allow on-the-fly switching of bitrate within the same sequence.

This implementation is mainly intended for a coding driven by the network condition or by the user preferences, avoiding the use of different pre-encoded and stored sequences. This approach is particularly suitable for real-time communications, provided that the feedback from the network or the user is immediate. We proposed the employment of an initially static table to select a suitable quantization parameter for each new bitrate request. This table has been obtained by performing different encodings of several sequences, and then observing the results. The proposed initial table can be computed by means of a recursive formula. During encoding, statistics on the already coded GOPs are used to update the table and so adjusting the values for the particular sequence content.

We demonstrated that this approach can achieve constant bitrate (CBR) coding with a higher precision than the standard encoder, being usually its error lower than 1%. Moreover, if information on per-GOP bitrate pattern is provided, the modified encoder can safely switch between the desired bitrates, fastly converging to the indicated value, so adapting nearly immediately to the new network available bandwidth or user preference with an error which can arrive up to 5% if the gaps between values we switch among are wide and the frames rate value is as high as 30 fps. Better results are obtained with lower fps and closer bitrate levels. This characteristic makes this implementation suitable for wireless communications, where low bitrates and fast network adaptivity are mandatory constraints for achieving satisfactory performance.

References

1. ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC: Advanced video coding for generic audiovisual services. ITU-T (2003)
2. Gallant, M., Kossentini, F.: Rate-distortion optimized layered coding with unequal error protection for robust internet video. IEEE Transactions on Circuits and Systems for Video Technology **11**(3) (2001) 357–372
3. Ahmed, T., Mehaoua, A., Boutaba, R., Iraqi, Y.: Adaptive packet video streaming over ip networks: a cross-layer approach. IEEE Journal on Selected Areas in Communications **23**(2) (2005) 385–401
4. Li, W.: Overview of fine granularity scalability in mpeg-4 video standard. IEEE Transactions on Circuits Systems Video Technology **11** (2001) 301–317
5. Radha, M.M., van der Schaar, M., Chen, Y.: The mpeg-4 fine-grained scalable video coding method for multimedia streaming over ip. IEEE Transactions on Multimedia **3** (2001) 53–68
6. Chang, S., Vetro, A.: Video adaptation: Concepts, technologies and open issues. IEEEProceedings of the IEEE **93**(1) (2005) 148–158
7. Radulovic, I., Frossard, P., Verscheure, O.: Adaptive video streaming in lossy networks: Versions or layers? In: Proc. IEEE Int. Conf. on Multimedia & Expo, Taipei, Taiwan (2004) 1915–1918
8. Li, B., Liu, J.: Multirate video multicast over the internet: An overview. IEEE Network, Special Issue on Multicasting: An Enabling Technology **17**(1) (2003) 24–29
9. Ortega, A., ed. In: Variable Bit-Rate Video Coding, in Compressed Video over Networks. M.-T. Sun and A. R. Reibman, Eds, New York, NY, USA (2000) 343–382
10. Kim, J., Wang, Y., Chang, S.: Content-adaptive utility-based video adaptation. In: Proc. IEEE Int. Conf. on Multimedia & Expo, Baltimore, Maryland (2003)
11. H. Song, C.K.: Rate control for low-bit-rate video via variable-encoding frame rates. IEEE Transactions on Circuits and Systems for Video Technology **11**(4) (2001) 512–521

User Authentication Based on JPEG2000 Images

Maria Teresa Carta, Barbara Podda, and Cristian Perra

DIEE, Department of Electrical and Electronic Engineering,
University of Cagliari,
Piazza D'Armi, Cagliari 09123, Italy
{mariateresa.carta, barbara.podda, cperra}@diee.unica.it

Abstract. The problem of user authentication is particularly interesting for increasing information security on Internet based application. The more used authentication systems are based on alphanumeric passwords and their main weak point is the difficulty to remember them. The paper aims on the study of a framework for user authentication which uses JPEG2000 images as passwords. This approach is based on the human remarkable ability to remember images. The proposed image based authentication (IBA) system avoids the common hacker attacks and it is usable in heterogeneous networks, it is also cost effective, and user-friendly. The system is developed for Internet based application using personal digital assistants or personal computers and it can be easily extended to others network applications.

1 Introduction

The Internet network offers many opportunities for industry, economy, technology etc. Due to the World Wide Web evolution, every day a quick exchange of information is possible from a part of the globe to another. This information exchange needs to be protected from intrusions of unauthorized subjects (for example during an economic transaction). The authentication is a procedure that permits to recognize the user which wants to access to the information. During this procedure an authorized part assesses the user identity with a specific test. In the proposed method the test is based on images recognition. Studies about cognitive process shown that the human brain can remember images more easily than passwords [3]. The common hacker attack techniques [2,4,5] are ineffective against the proposed framework, because they are performed for system based on alphanumeric strings or on biometric features. The project is developed within the JPEG standardization activities related to the topic of image based access systems and it is based on JPEG2000 standard and JPEG2000 Interactive Protocol standard.

2 The Framework

In the image based authentication (IBA) method proposed, a personal image and an user ID is required in order to authenticate the user. A generic IBA system should collect user images to be used by a challenge and response protocol for identifying the

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 207 – 213, 2006.
© Springer-Verlag Berlin Heidelberg 2006

user. The whole authentication procedure is based on the recognition of an altered image portion (i.e.: challenge) and, successively on the transmission to the server of the original image portion (i.e.: response).

In order to make feasible the image exchange and the user authentication process, some functionalities, such as scalability, progressive image transmission, client/server interactivity, are necessary. These functionalities are provided in the emerging JPEG2000 standard for image coding published by the JPEG committee (ISO/IEC JTC 1/SC 29/WG 1).

The IBA architecture consist of an IBA Server and an IBA Client as shown in Fig. 1.

In the system proposed, all the computational complexity is concentrated in the IBA server and consequently the client implementation is very simple to realize. Furthermore the information exchange between client and server need to be protected, and for this reason the Client-Server communications is performed on a secure channel.

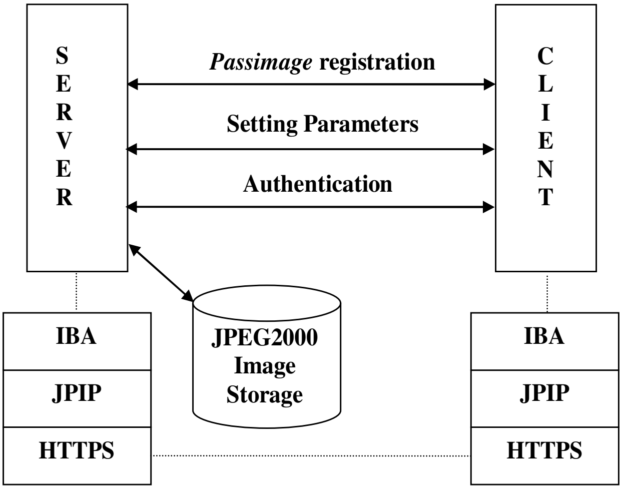


Fig. 1. The proposed Image Based Authentication (IBA) framework

2.1 Registration

In the first step of the method, the user registers him-self through the submission of an user ID and of a personal image. The registration is divided into four steps represented in Fig. 2 and explained in this paragraph.

I. ***Registration Request*:** the user send his personal information: name, e-mail, personal user ID, etc. by using a simple web interface. In the same registration phase, he finally uploads his personal image (Passimage) to the server.

II. ***Image Processing*:** the server codes the image using JPEG2000 standard. The image is subdivided into non-overlapping tiles. This procedure allows to reduce memory requirements and, at the same time,. to decode only the selected parts of the image in opposition to the whole image decoding. The coded stream is then reordered in a stream media type suitable for subsequent JPIP connection. At this point the code-stream is reordered and completed with an index in order to have a fast recovery of the information packages. The image processed is called Passimage. The whole image processing procedure is completely presented in Fig. 3.

III. ***Storing Data*:** the server stores the user ID, the user information and the Passimage in a database.

IV. ***Notification*:** the server sends an e-mail to the user for confirming the registration. This step is performed in order to have a secure feed-back of the whole operation.

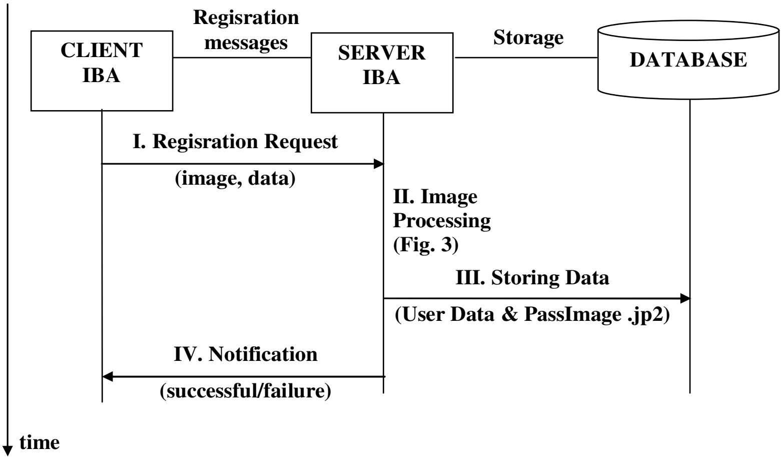


Fig. 2. Registration Process

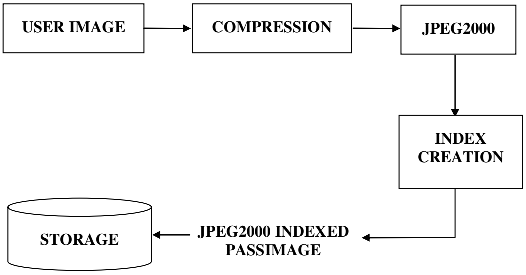


Fig. 3. Image Processing Procedure

2.2 Authentication

The authentication step is based on the main steps described in Fig. 4 which are completely explained in the rest of this paragraph.

I. ***Authentication Request*:** in order to allow the communication between user and server, the authentication system is characterized by a simple web interface. The user inserts his user ID in the web interface and he specifies the authentication modality by choosing it between two different possibility. In the first case the user has his personal image stored in the client or in a personal mobile device. In the second case the user does not have his Passimage but he can, however, authenticate him-self by using his personal image stored in the server database. The database has a large number of images that are organized in categories (e.g.: landscapes, animals, group of peoples, single person, object, etc.). This structure helps the user to find rapidly his Passimage. The following description is limited to the first case mentioned above, where the user has his Passimage.

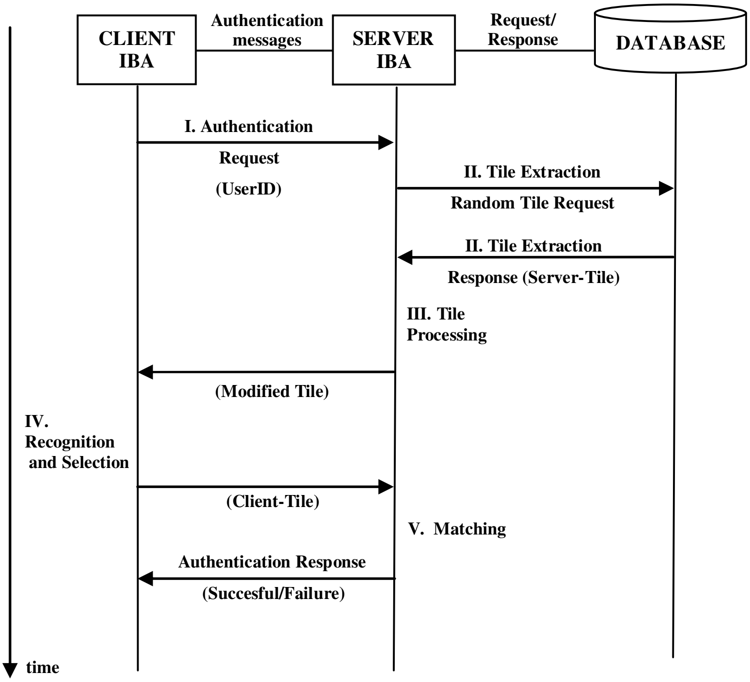


Fig. 4. Authentication Procedure

II. ***Random Tile Extraction*****:** the server identifies the user by his user ID. Each user ID has a corresponding Passimage known by the server. At this point, the server extracts randomly a tile (server-tile) from the code-stream associated to a specific Passimage. This random access is realized by exploiting the functionalities of the JPIP standard. The tile (server-tile) is selected through a pair of random numbers that are generated from a specific function. The Passimage is characterized by some JPIP parameters as: tile resize (rsizeX, rsizeY); tile fsize (fsizeX, fsizeY) that are chosen from the server on the basis of the client type image resolution (e.g.: PC, PDA, etc.). The tile offset (oX, oY) is another JPIP parameter that identifies the position of the tile in the image and it is different for every authentication. The JPIP parameters are shown in Fig. 5.

III. ***Tile Processing:*** the server processes the server-tile in order to send it through the channel quickly and in a secure modality. The first processing step is the tile conversion to an uncompressed format. For avoiding to send the original tile trough the channel the tile is altered using a generic filter (e.g.: median filter) that maintains the tile recognizable but at the same time introduces some alterations to the original image portion (server-tile). Finally, the obtained tile is compressed in .jp2 format and sent to the client. The JPIP standard allows the delivery of portion of JPEG2000 images in arbitrary order. JPIP defines an interactive protocol to achieve an efficient exchange of JPEG2000 images and related data.

IV. ***Recognition and Selection*****:** the user can observe through a graphical interface (e.g.: a webpage) both the tile received from the server and the original Passimage that is stored in the client or in an external storage device (e.g.: USB memory pen, floppy disk, etc). The user recognizes the received tile, and after having selected the corresponding region in the original image he sends it to server.

V. ***Matching*****:** when the server receives the client-tile it matches each of its pixels with the corresponding pixels of the server-tile. If the measure of the MSE (Mean Square Error) is equal to zero the user is authenticated. In others words the method proposed allows the user authentication only when the correct original tile is sent to the server.

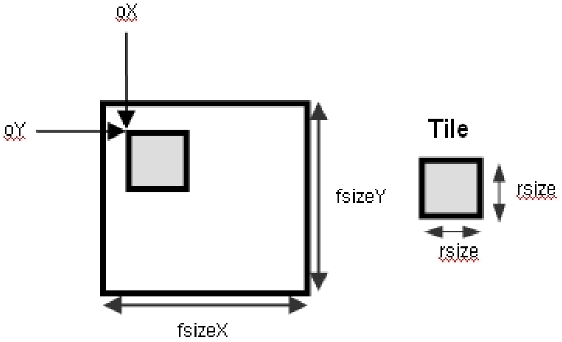


Fig. 5. JPIP parameters used for the random Tile Extraction. Tile resize (rsizeX, rsizeY); tile fsize (fsizeX, fsizeY), tile offset (oX, oY).

2.3 The Passimage

The user could have some difficulties in the recognition step if he chooses a Passimage characterized by tiles having uniform grey-levels, or tiles that are very similar each others (little variable background, e.g.: sky, sea, etc.). In order to avoid this problem it has been introduced a function that executes an entropic control on the image.

In the registration phase the server executes an entropic control on the whole image, consequently a very little variable image can not be accepted from the server as a valid Passimage.

Sometimes, the images have an high variability but they present some uniform tiles. In order to overcome this problem, in the authentication phase each random server-tile, before being processed, is controlled by a function that measures its entropic value. If the tile does not exceed a fixed entropic threshold, the server proceeds iteratively with a new random selection until the tile reaches the established entropic threshold. Due to this entropic control operation, the authentication phase can have a variable duration, sometimes longer respect to the best case, but, at the same time, the system reliability is increased.

3 Considerations

The image based authentication method proposed solves many problems that affect the traditional authentication systems, in particular it overcomes the problems related to the difficulty in remembering passwords as required by the common internet applications. The common password or PIN based systems requires very frequently to the user to change his password in order to increase the security level. This operation can be definitively simplified if the system is based on personal pictures, because the user does not have the problem to remember the new password but he has only the task to recognize an image that he well knows. This fact encourages the user to change frequently the personal Passimage, by guaranteeing an higher security level to the system.

The whole image is transmitted in the channel only during the registration phase. It is not possible to perform an observation attack, because in every authentication the tile sent to the user is different and because the tile sent to the client is opportunely modified before being sent to the user. The JPEG2000 standard advantages have been exploited in order to reduce the server memory occupation, and at the same time in order to have an efficient transmission of the tile above the channel. By considering the more common communication channels, it has been evaluated the time necessary to perform the authentication step and Table 1 shows the results obtained.

The presented results can be considered very satisfactory for the intended application of this framework.

The results have shown that the registration and the authentication steps have been easily executed without lost of time and without errors. Each test participant easily assimilated the correct use of the system and expressed agreement for the new IBA system.

In opposition to the others authentication systems (e.g. credit cards based method, biometric systems, etc.), the IBA system proposed does not require the use of specific hardware devices and for this reason it is very cheap to realize. At the same time the software that implements the method has a low computational cost.

Table 1. Time of tile transmission in the more common communication channels expressed in milliseconds

File.j2k	Analogical				GPRS		
[Byte]	Modem	ISDN	ADSL	GSM	theoretic	effective	UMTS
1650	293	206	21	1375	77	440	155
1591	283	199	20	1325	74	424	150
1931	343	241	24	1609	90	515	182
1127	200	141	14	939	53	301	106

In comparison to the most modern techniques based on the IBA, where the images are used only in the user interface, the system demonstrated to be completely innovative, because the whole authentication procedure is based on the image recognition. This fact is made possible from two new standards: JPEG2000 for the image compression and JPIP (JPEG2000 Interactive Protocol) for the interactive exchange of images and data through the network.

4 Considerations

The image based authentication system proposed in this paper has been developed to be a valid alternative to the conventional systems and to have a robust architecture against the common internet attacks. This framework has been developed to overcome the weak points of the conventional systems and the results have shown its effectiveness. In particular, the IBA system provides a new approach that overcomes the problems related to the human effort in remembering one or more passwords as required by the common internet applications. The main element of system is the human capability to recognize immediately an image. The user authentication is obtained by recognizing a tile randomly extracted by a specific personal image (Passimage). JPEG2000 and JPIP standards assure a minimum allocation of memory server and communication channel. The IBA architecture is simple, user-friendly, cost-effective and it can be easily adapted to different kind of applications.

References

1. Ebrahimi, T.: JPEG 2000 new activities and explorations. ISO/IEC JTC 1/SC 29/WG1N3170, December (2003)
2. Pering, T., Sundar, M., Light, J., Want, R.: Photographic Authentication through Untrusted Terminals. IEEE Pervasive Computing, January (2003) 30-36.
3. Paivio, A., Rogers, T. B., Smythe, P. C.: Why are pictures easier to recall than words? Psychonomic Science, 11(4), (1968) 137-138
4. Taada, T., Koike, H.: Awase-E: Image-based Authentication for Mobile Phones using User's Favorite Images. Proceedings on MobileHCI, (2003)
5. Jansen, W., Gavrila, S., Korolev, V., Ayers, R., Swanstrom, R., Picture Password: A Visual Login Technique for Mobile Devices, National Institute of Standard and Technology. NISTIR7030, July (2003)
6. Jermyn, I., Mayer, A., Monrose, F., Reiter, M.K., Rubin, A. D.: The Design and Analysis of Graphical Passwords. Proceedings of the 8th USENIX Security Symposium, August, Washington DC (1999)
7. Dhamija, R., Perrig, A., Deja Vu: A User Study Using Images for Authentication. 9th Usenix Security Symposium, August (2000) 45-48
8. Prandolini, R., Houchin, S., Colyer G.: JPEG 2000 image coding system – Part 9: Interactivity tools, APIs and protocols – Final Committee Draft 2.0. ISO/IEC JTC 1/SC 29/WG1N3174, December (2003)
9. Perra, C., Giusto, D.D.: A Framework For Image Based Authentication. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA, March 19-23 (2005)

Evaluation of Color Image Segmentation Algorithms Based on Histogram Thresholding

Patrick Ndjiki-Nya, Ghislain Simo, and Thomas Wiegand

Image Communication Group, Image Processing Department,
Fraunhofer Heinrich-Hertz-Institut,
Einsteinufer 37, Berlin 10587, Germany
{ndjiki, stueber, wiegand}@hhi.de

Abstract. Image segmentation is an essential processing step in texture analysis systems, as its accuracy has a significant impact on the quality of the final analysis result. The downside of texture analysis is that segmentation is one of the most difficult tasks in image processing. In this paper, algorithms for improved color image segmentation are presented. They are all based on a histogram thresholding approach that was developed for monochrome images for it has proven to be very effective. Improvements over the genuine segmentation approach are measured and the best optimization algorithm is determined.

1 Introduction

Texture analysis typically requires segmenting an image or video sequence into uniformly textured regions. The segmentation step is both critical and essential, as its accuracy has a significant impact on the quality of the final analysis result. However, image segmentation is also one of the most difficult tasks in image processing.

The complexity of the texture analysis task mainly arises from two problems [1]. The first difficulty lies in the inherent variability of natural textures, which has implications for any local texture measure used in the context of texture analysis. In fact, region boundaries must be identified without making too strong and thus unrealistic assumptions with regard to the homogeneity of the texture within a given region. The second difficulty is called uncertainty problem and can be described as follows. To achieve a detailed representation of gray value discontinuities, high spatial resolution is required. However, given the latter input, boundary detection methods are likely to generate spurious responses within textured regions. On the other hand, in order to segment textured images into meaningful regions, a sufficiently large area averaging process is needed to reduce the fluctuations in texture properties. Although increasing texture regions' homogeneity, this lowpass operation yields blurred boundaries. Obviously, the uncertainty problem arises from the conflict between the simultaneous measurement of texture properties and corresponding spatial location. This chicken and egg dilemma is however a central issue in segmentation applications.

Many segmentation methods are based either on local pixel similarities or corresponding discontinuities. Similarity or homogeneity-driven approaches encompass

L. Atzori et al. (Eds.): VLBV 2005, LNCS 3893, pp. 214 – 222, 2006.
© Springer-Verlag Berlin Heidelberg 2006

thresholding, clustering, region growing, as well as splitting and merging. Discontinuity or boundary-based approaches partition images based on criteria as edges. Gray value gradients are assumed to be smooth within objects and steep at object boundaries in this framework. The smoothness criterion should account for the inherent randomness of natural textures in order to avoid unrealistic homogeneity assumptions about within-object textures.

Freixenet et al. [2] have evaluated several representative spatial image segmentation approaches. They define test conditions comprising natural and synthetic images. The segmentation results are evaluated based on objective measures introduced by Huang and Dom [3]. These measures allow a precise evaluation of boundary and region-related performance of clustering algorithms. The experimental results obtained in [2] show that, in general, the multiresolution algorithm proposed in [4] yields the best results, both in terms of segmentation accuracy and computational complexity. Thus a similar algorithm [5] to the one described in [4] is used as a basis for the spatial texture analysis in this work.

The remainder of the paper is organized as follows. In Sec. 2, the baseline approach [5] is described in-depth. The optimization algorithms are presented in Sec. 3 and 4. The objective quality assessment measures used in this paper are presented in Sec. 5. Finally, the experimental results are shown in Sec. 6.

2 Multiresolution Approach by Spann and Wilson

In [5] and [4], the randomness and uncertainty problems, that are key issues in segmentation as explained above, are tackled by using a multiresolution analysis approach. The fundamental assumption of these approaches is the invariance of object properties across scales. The segmentation algorithm developed by Spann and Wilson [5] will be presented in detail in the following.

Spann and Wilson's approach generates a multiresolution image pyramid by applying a quad-tree smoothing operation on the original image. Homogeneous regions are extracted at the highest level of the quad-tree, i.e. at the lowest spatial resolution, via statistical classification. The latter is based on a local centroid algorithm described in [6]. The classification step is followed by a coarse to fine boundary estimation based on the partition obtained at the top level of the pyramid. No a priori information, such as the number of segments, is required in this framework.

2.1 Quad-Tree Smoothing

The smoothing operation is the first step of the algorithm as presented in [5]. It may be operated on the original image or a transformed representation of it [5]. The smoothing is performed using a conventional quad-tree approach [5], which yields a hierarchy of lowpass filtered versions of the original image, such that successive ascending levels correspond to lower frequencies.

Not only is quad-tree smoothing a fast operation, it also allows to balance the reduction in measurement noise via smoothing against the bias related to the fusion of information from possibly distinct homogeneous regions. Given the highest possible pyramid level L of the considered image, the maximum smoothing gain can be obtained by truncating the quad-tree at level $\ell' < L$, while maintaining sufficient resolution to ensure accurate segmentation of all regions whose radius r fulfills the condition: $r \geq 2^{\ell'+1}$.

2.2 Local Centroid Clustering

Local centroid clustering is applied at the highest pyramid level to achieve consistent homogeneous texture regions. The algorithm described in [6] is used. It allows a non-parametric classification of the image plane with the lowest resolution as the basis of its gray value statistics. The local centroid algorithm iteratively moves the bin populations of the original histogram $h(i)$ to their center of gravity within a sliding window of size $2m+1$. Once the algorithm has converged, the classification is in principle achieved by mapping each bin of the original histogram to the corresponding center of gravity.

The implicit assumption of the local centroid method is that images are composed of regions with different gray level ranges, such that the corresponding histogram can be separated into a number of peaks or modes, each corresponding to one region, and there exists a threshold value corresponding to the valley between two adjacent peaks [7]. Thus the number of classes obtained through local centroid clustering depends on the "peakiness" of $h(i)$ and on the window size $2m+1$.

2.3 Boundary Estimation

The uncertainty problem introduced above must now be dealt with in order to robustly extract reliable region contour information from the image data. Spann and Wilson remove uncertainty by introducing a fundamental assumption: The region properties' invariance across the scales spanned by the quad-tree. Given this hypothesis, classification results obtained for image planes at higher pyramid levels can be propagated to lower levels of the tree to initialize the new classification cycle. Non-boundary pixels are assigned to the same class across levels, while boundary nodes

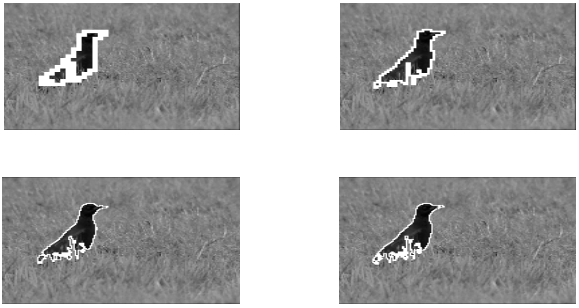


Fig. 1. Coarse to fine boundary refinement by Spann and Wilson [5]. Image planes are depicted in the order 3 (top left), 2 (top right), 1 (bottom left), and 0 (bottom right).

are re-classified at the lower pyramid level in such a way that the boundary width is reduced by a factor of two on each step down the quad-tree. This is illustrated in Fig. 1, where four pyramid levels are considered ($L = 3$), the original image being at level 0 and the lowest spatial resolution at level 3. Note that image planes at higher pyramid levels have been interpolated to achieve the size of the original image for better visualization of the boundary refinement (cp. Fig. 1).

Histogram thresholding approaches have been first introduced for monochrome images and widely used for segmentation [8]. Hence, they do not take advantage of the information available in the typically three color channels. In this work, approaches have been developed that alleviate the above-mentioned drawback. The latter will be presented in-depth in the following sections.

3 Histogram Thresholding Via Color Channel Pruning

The channel pruning approach consists in selecting a color channel for multiresolution segmentation, such that the within-cluster variance is minimized and the between-cluster distance is maximized. Algorithms following a similar idea as can be found in [7]. The formal definition of the cost function specific to this work is given by

$$\mathrm{E} = -\frac{N_C\left(N_C - 1\right)}{2} \frac{\sum\limits_{i \in C} \sum\limits_{\substack{j \in C - \{i\} \\ j > i}} \left\| i - j \right\|}{\sum\limits_{i \in C} \sigma_i^2} \tag{1}$$

where C represents the set of cluster centroids sorted in ascending order. N_C is assumed to be the number of cluster centroids in C. Note that the centers of gravity correspond to bin locations within a given histogram. The quotient depending on N_C is a normalization factor to the numerator (sum of inter-cluster distances) of the second quotient. Note that the norm used to determine the distance between two clusters can be chosen according to application requirements. However, the ℓ_1 norm is used in this work. σ_i^2 is assumed to be the within-cluster variance of the cluster represented by the centroid i. The histogram of the input color channel that minimizes the cost function is used for multiresolution segmentation, as it potentially yields the best result given the adopted segmentation strategy. Channel discrimination is done at the highest pyramid level, i.e. at the lowest spatial resolution, in order to achieve robust homogeneous regions.

The required cluster centroids as well as limits are extracted using the local centroid clustering algorithm described in [6].

4 Histogram Thresholding Via Redundance Elimination

Although the approach presented in the previous section provides automatic selection of the best of the available color channels for the segmentation task, it does not take the correlation between the color components into account. Hence the segmented regions are based on just one of at least three color features.

The redundancy elimination approach presented in this section is similar to the algorithm by Ohta et al. [9]. Color features with high discrimination power are extracted using principal component analysis (PCA) [10]. PCA is an approach that is typically used for linear dimensionality reduction and that inherently allows for explicit control of the error introduced by dimensionality reduction.

Consider a region $\mathcal{S}$ in a D-dimensional color space. RGB [7] will be considered in the following ($D = 3$), but other color spaces (e.g. HSV [7]) may also be used. Let the distributions of R, G and B in $\mathcal{S}$ be h_{R}, h_{G}, and h_{B}, respectively. Let Σ represent the covariance matrix of the set of row vectors of the matrix $[h_{\mathrm{R}}\ h_{\mathrm{G}}\ h_{\mathrm{B}}]$ of dimension Nx3, where N is the histogram resolution. Further assume that $\lambda_1 \geq \lambda_2 \geq \lambda_3$, where λ_i are the eigenvalues of Σ. Then the color features $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ defined as

$$\mathcal{F}_i = \mathrm{f}_{\mathrm{Ri}} \times \mathrm{R} + \mathrm{f}_{\mathrm{Gi}} \times \mathrm{G} + \mathrm{f}_{\mathrm{Bi}} \times B \tag{2}$$

where the eigenvectors of Σ that are given by $f_{\mathrm{i}} = (\mathrm{f}_{\mathrm{Ri}}, \mathrm{f}_{\mathrm{Gi}}, \mathrm{f}_{\mathrm{Bi}})$ (RGB color space) can be shown to be uncorrelated with $\mathcal{F}_1$ having the largest variance equal to λ_1, and thus the largest discriminant power. $\mathcal{F}_2$ has the largest discriminant power among the vectors orthogonal to $\mathcal{F}_1$ [10]. In this work, $\mathcal{F}_1$ is used for multiresolution segmentation.

5 Objective Evaluation of Segmentation Masks

5.1 Huang and Dom's Measures

Segmentation results in this work are evaluated using the methods proposed by Huang and Dom [3]. The automatic segmentation evaluation is split into two parts. The first one consists in a boundary-based evaluation, while the second one corresponds to a region-based quality assessment approach. Both parameter classes will be presented into detail in the following.

Weighted Boundary Error Rates

Boundary error rates basically measure the discrepancy between true and segmented region contours. Assume that $\mathcal{G}_{\mathbf{B}}$ is the set of true boundaries. Let $\mathcal{S}_{\mathrm{B}}^{\mathrm{TP}}$ and $\mathcal{S}_{\mathrm{B}}^{\mathrm{FP}}$ be the set of true and false positive contour pixels respectively. The missing (e_B^m) and the false boundary (e_B^f) rates can be derived from above definitions as follows

$$e_B^m = \frac{\left|\mathcal{G}_{\mathrm{B}}\right| - \left|\mathcal{S}_{\mathrm{B}}^{\mathrm{TP}}\right|}{\left|\mathcal{G}_{\mathrm{B}}\right|}\ ,\ e_B^f = \frac{\left|\mathcal{S}_{\mathrm{B}}^{\mathrm{FP}}\right|}{\left|\mathcal{G}_{\mathrm{B}}\right|} \tag{3}$$

where $\left|\ \right|$ denotes the size of a set. e_B^f can be seen as a false positives rate r_B^{FP}, while the true positives rate r_B^{TP} can be derived from e_B^m as follows

$$r_B^{TP} = 1 - e_B^m = \frac{\left|\mathcal{S}_{\mathrm{B}}^{\mathrm{TP}}\right|}{\left|\mathcal{G}_{\mathrm{B}}\right|} \tag{4}$$

The "closeness" of the location of the true and the segmented region contours is determined by the missing (w_B^m) and the false (w_B^f) boundary weights, where the weights represent the mean distance between the misclassified samples and the ground truth boundary. The distance between a point and a set is thereby defined as the minimum absolute distance from the point to all points of the considered set. w_B^m and w_B^f are normalized with the length of the main diagonal of the considered image and thus lie within the interval 0.0 to 1.0.

Region Error Rates

Region error rates measure the discrepancy between true and segmented regions. This is done by applying the directional Hamming distance [3] $d_H(\mathcal{S}_1 \rightarrow \mathcal{S}_2)$ from a segmentation $\mathcal{S}_1 = \{\mathcal{R}_{11}, \mathcal{R}_{12}, \mathcal{R}_{13}, \ldots\}$ to another segmentation $\mathcal{S}_2 = \{\mathcal{R}_{21}, \mathcal{R}_{22}, \mathcal{R}_{23}, \ldots\}$. For that, mapping of the regions of $\mathcal{S}_2$ onto those of $\mathcal{S}_1$ is required and realized such that the region $\mathcal{R}_{2j}$ is associated with the region $\mathcal{R}_{1i}$ if and only if $\mathcal{R}_{2j} \cap \mathcal{R}_{1i}$ is maximal. The directional Hamming distance then measures the total area for which $\mathcal{R}_{2j} \cap \mathcal{R}_{1k}$ $(k \neq i)$ is non-maximal.

$$e_R^m = \frac{d_H\left(\mathcal{S}_\mathrm{R} \rightarrow \mathcal{G}_\mathrm{R}\right)}{\left|\mathcal{S}_\mathrm{R}\right|} \quad , \; e_R^f = \frac{d_H\left(\mathcal{G}_\mathrm{R} \rightarrow \mathcal{S}_\mathrm{R}\right)}{\left|\mathcal{S}_\mathrm{R}\right|} \tag{5}$$

Given the definition of the Hamming distance, Huang and Dom [3] define a missing and a false alarm rate denoted e_R^m and e_R^f respectively. It is assumed that $\mathcal{G}_\mathrm{R}$ is the set of true regions, while the set of segmented regions is called $\mathcal{S}_\mathrm{R}$. e_R^m then measures the percentage of the samples in $\mathcal{G}_\mathrm{R}$ being mistakenly segmented into wrong regions in $\mathcal{S}_\mathrm{R}$, while e_R^f describes the percentage of image samples in $\mathcal{S}_\mathrm{R}$ falling into regions of $\mathcal{G}_\mathrm{R}$ that are non-maximal intersected with the region under consideration. The formalization of the missing and the false alarm error rates is given in (5).

$$r_R^{TP} = 1 - e_R^m \tag{6}$$

Similarly to the boundary error rates, e_R^f can be seen as a false positives rate r_R^{FP}, while the true positives rate r_R^{TP} can be derived from e_R^m as shown in (6).

5.2 Receiver Operating Characteristic Curve

Receiver Operating Characteristic (ROC) curves were developed in the 1950's in the field of radio signal processing [11]. We use ROC curves in the context of image segmentation to evaluate the accuracy of a given segmentation approach given a ground truth set, i.e. the ability of the segmentation algorithm to provide accurate image partitions. The abscissa of the ROC curve represents the false positive (FP) rate, while the ordinate corresponds to the true positive (TP) rate. (FP,TP) pairs are measured for several configurations of the segmentation algorithm, each pair corresponding to the average performance of the considered approach over the ground truth set. The area under the ROC curve, also called AUC, can be interpreted as the percentage of randomly drawn data pairs (one from the true positive and one from false positive class) for which the segmentation algorithm yields a correct classification. The accuracy of the segmentation approach is proportional to AUC. The ideal AUC corresponds to 1.0, i.e. 100% TP independently of the FP rate. An AUC of 50% stands for a worthless segmentation approach, while 70%-80% are obtained for a fair, 80%-90% for a good, and 90%-100% for a very good algorithm.

6 Experimental Results

The evaluation of the segmentation algorithms presented above is conducted with a ground truth set of 100 images from the Corel Gallery™ (US version, 07/1998) database. The images are selected in consideration of the lighting conditions, the presence/absence of details in the images and a "good" coverage of the HSV color space. A further discrimination criterion consists in selecting only images that show a large variance in more than one dimension in the HSV color space. This is an important prerequisite to ensure meaningful results with regard to the best optimization algorithm.

For each ground truth image, a reference partition is generated manually. The clusters thereby reflect a semantic decomposition of the scene. Each ground truth image is further segmented using a segmentation approach at a given configuration and the obtained partition is compared to the reference partition using the measures defined in Sec. 5.1. Each of the segmentation algorithms presented in Sec. 2, 3, and 4 feature two degrees of freedom, i.e. the window size m and the number of pyramid levels ℓ' (cp. Sec. 2).

Our experiments show that ℓ' can be set to four without penalizing any of the ground truth images. Moreover, it was found that m has the most significant impact

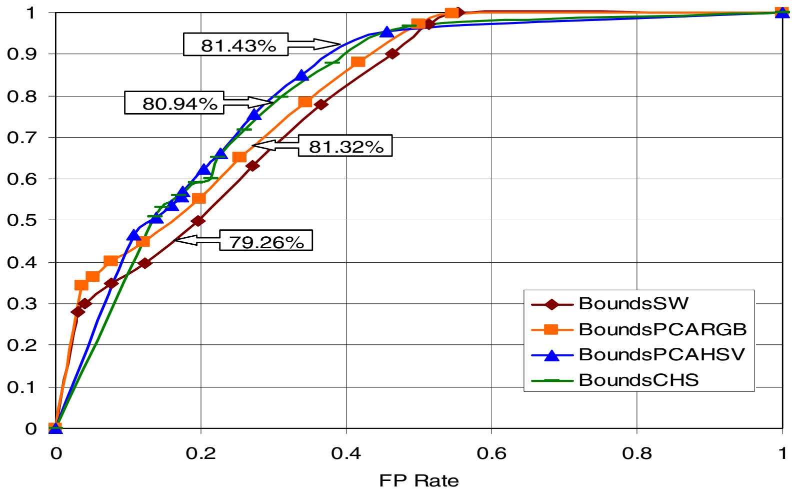


Fig. 2. ROC curves for boundary error rates

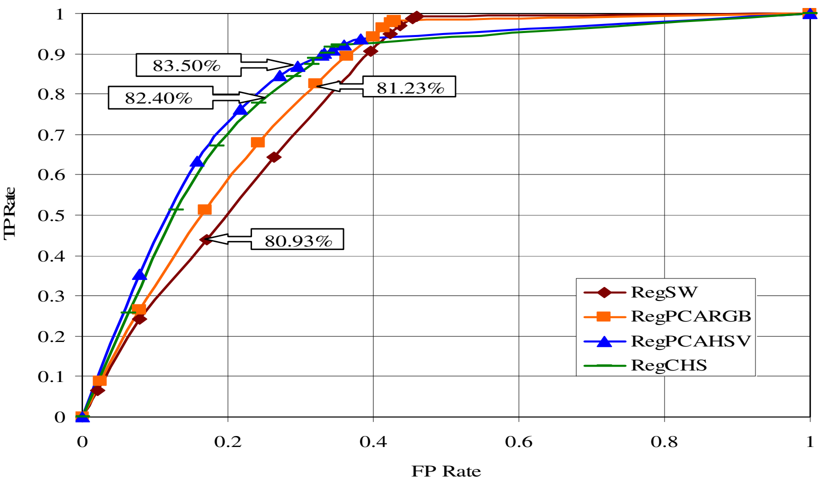


Fig. 3. ROC curves for region error rates

on segmentation quality compared to ℓ'. For that, m is drawn from the set $\{1, 5, 10, 15, 20, 25, 30, 35, 40, 45\}$ in our experiments. The results of our evaluations are depicted in Fig. 2 and Fig. 3 for boundary and cluster quality assessment respectively. The AUC values are given for each optimization algorithm in both figures.

It can be seen that better results than the genuine segmentation approach by Spann and Wilson [5] are achieved for each of the optimization algorithms. This shows that the algorithms proposed in this paper are more powerful than [5] (SW) for color images as expected. Boundary accuracy is improved from fair to good compared to [5] in all cases. w_B^m is significantly reduced for all optimization approaches, both in terms of mean (e.g. from 0.05 to 0.025 for CHS) and variance (e.g. from 0.22 to 0.1 for CHS).

The mean w_B^f is almost the same for all optimization approaches compared to [5] (0.025), but lower variances are measured (e.g. reduction from 0.11 to 0.075 for PCAHSV). PCAHSV yields the best overall results. An exemplary segmentation result is depicted in Fig. 4.

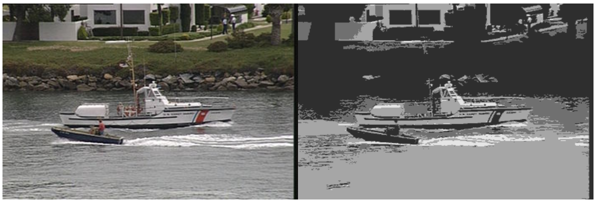


Fig. 4. Segmentation result obtained with the redundancy elimination algorithm (HSV)

7 Conclusions

Segmentation approaches for improved color image segmentation are presented in this paper. They are all based on a histogram threshold approach that is one of the best algorithms for monochrome image segmentation. It could be shown that, by exploiting the multi-feature nature of color images, the optimization algorithms yield better results than the approach in [5] as expected. Furthermore, it is found that the redundancy elimination approach with HSV yields the best results.

References

1. Wilson, R., and Spann, M.: Image Segmentation and Uncertainty. Pattern Recognition and Image Processing Series, Research Studies Press Ltd, England (1988)
2. Freixenet, J., et al.: Yet Another Survey on Image Segmentation: Region and Boundary Information Integration. Proc. 7th European Conference on Computer Vision-Part III (2002) 408-422
3. Huang, Q., and Dom, B.: Quantitative Methods of Evaluating Image Segmentation. Proc. ICIP, Vol. 3 (1995) 53-56
4. Wilson, R., and Spann, M.: Finite Prolate Spheroidal Sequences and their Applications II: Image Feature Description and Segmentation. IEEE Trans. on PAMI, Vol. 10 (1988) 193-203
5. Spann, M., and Wilson, R.: A Quad-tree Approach to Image Segmentation which Combines Statistical and Spatial Information. Pattern Recognition, Vol 18, Nos. 3/4 (1985) 257-269
6. Wilson, R., Knutsson, H., and Granlund, G. H.: The Operational Definition of the Position of Line and Edge. Proc. ICPR (1982)
7. Cheng, H. D., Jiang, X. H., Sun, Y., and Jingli Wang: Color Image Segmentation: Advances and Prospects. Pattern Recognition, Vol. 34 (2001) 2259-2281
8. Littmann, E., and Ritter, H.: Adaptive Color Segmentation – A Comparison of Neural and Statistical Methods. IEEE Trans. on Neural Networks, Vol. 8, No. 1 (1997) 175-185
9. Ohta, Y., Kanade, T., and Sakai, T.: Color Information for Region Segmentation. Computer Graphics and Image Processing, Vol. 13, No. 3 (1980) 222-241
10. Bishop, C. M.: Neural Networks for Pattern Recognition. Oxford University Press (1995)
11. Hanley, J. A., and Mc Neil, B. J.: The Meaning and Use of the Area under the Receiver Operating Characteristic (ROC) Curve. Radiology, Vol. 1, No. 143 (1982) 29-36

Author Index