A LOW BIT RATE SEGMENTED VIDEO CODEC WITH HYBRID MOTION ESTIMATION AND INHERENT BIT RATE CONTROL CAPABILITY

V. A. Christopoulos, J. Cornelis

Vrije Universiteit Brussel, VUB-ETRO (IRIS), Pleinlaan 2, 1050 Brussels, Belgium
Tel: +32 2 6292982; fax: +32 2 6292883
e-mail: {vschrist,jpcornel}@etro.vub.ac.be

ABSTRACT

In this paper a segmented video codec with hybrid motion estimation and inherent bit rate control capability is presented. The first frame in the data is always encoded in intraframe mode, while the rest of the frames are encoded in interframe mode. The interframe encoding is based on (1) hybrid conventional/perspective block motion vector estimation and coding, and (2) coding of the prediction error (Displaced Frame Difference, "DFD") using segmented image coding techniques.

We present a way to partition the DFD in moving and static regions and we explain how this classification strategy can be used as a means to control the bit rate. The simulation results show that the hybrid motion estimation technique outperforms the conventional full search block-matching method by improving the overall PSNR for reduced bit rate, and that the bit rate control strategy, although simple, is very efficient for monitoring both the bit rate and the quality degradation.

1. INTRODUCTION

Interactive multimedia communications, videophone, tele-surveillance, tele-control, remote consultation of multimedia databases, mobile audio-visual communications, etc., are examples of services and applications which can only be brought into wide spread usage if they can operate via low bit rate channels. *Segmented image coding (SIC) techniques* [6], which have already shown their maturity in still image compression at high compression ratios (>100 for colour images), are currently also applied in video compression.

In contrast to the well-known block transform coding schemes (e.g. JPEG for still images), SIC segments the image into objects (non-rectangular regions), and codes their contours (e.g. using chain codes) and their internal texture[1] (e.g. approximating it by a linear combination of orthonormal basis functions). SIC is particularly useful for very low bit rate applications because, at high compression ratios (CRs), it yields a better subjective image quality compared to that obtained by the conventional block based coding schemes (JPEG, MPEGx, H26x), since the objectionable blocking artefact is avoided and the image quality degrades more gently with decreasing bit rate. This asset of SIC together with its ability to handle content based functionalities, are two of the most important reasons for the attention paid to SIC techniques within the MPEG4 framework.

Much of the temporal correlation between frames of an image sequence can be removed by using motion compensation to predict each frame and then coding only the prediction error (Displaced Frame Difference, "DFD") and the motion field. The conventional approach of performing motion compensation, as defined in the H26x and MPEGx standards, relies on the assumption that the motion between the two frames is of a translational nature. This approach has been widely adopted, because it is of a simple form and does not have significant computational complexity. However, it causes problems in scenes with multiple moving objects, rotating objects and zoom, giving rise to a number of serious problems (e.g., block-effects in motion compensated predicted images, physically meaningless estimated motion vectors, etc.), thus making the current video coding standards inadequate for low bit rate coding. To develop a more effective motion compensation technique for low bit rate video coding, we have adopted a more sophisticated spatial transformation, the perspective transformation, and we have developed a motion estimation algorithm suitable for this transformation. We have combined this new motion estimation algorithm and the conventional block-matching motion estimation algorithm into a hybrid scheme capable of handling complex movements while the complexity of the algorithm is kept low. We have integrated this hybrid motion estimation algorithm into a SIC video codec [2][4], and its influence on both bit rate and quality has been studied. The results indicate an improvement in quality for decreased bit rate, compared to the conventional full search block-matching motion estimation technique.

Video coding and transmission schemes should be capable of monitoring the bit rate and ensuring the best level of quality under a given transmission channel's bandwidth. In this paper, we show that the classification strategy used in [2][4] to partition the DFD in moving and static regions can be used for bit rate control. We present simulation results which demonstrate the efficiency of this control strategy to monitor the bit rate and the quality degradation.

The paper is organised as follows. In section 2, the hybrid motion estimation algorithm and the bit rate control strategy are presented. Some simulation results obtained for typical colour videophone sequences are given in section 3. Finally, some conclusions and hints for further research are drawn in section 4.

[1] Texture is a generic term that refers to any quantitative colour/luminance property of the pixels in the objects.

2. THE SEGMENTED VIDEO CODEC

The first frame in the data is always encoded in intraframe mode, while the rest of the frames are encoded in interframe mode. The interframe coding procedure consists of the following steps (refer to [2][4] for a block diagram of the encoder):

- *Hybrid motion estimation*
- *Segmentation of the DFD*
- *Separation of regions into foreground/background (moving/static)*
- *Contour (simplification and) coding*
- *Reconstruction*

In the following section we briefly describe the hybrid motion estimation technique and we show that the way regions in DFD are classified into *moving* and *static* can be used as a simple means to control the video bit rate. For a description of the rest of the coding components we refer to [2][4].

2.1 The hybrid motion estimation technique

Block matching motion compensation (BMMC) algorithms can be defined as techniques that divide the image into blocks and estimate a set of motion parameters for each block. The predicted image of the n_{th} frame $\hat{I}_n(x,y)$ is then obtained from the decoded image of the previous frame $\tilde{I}_{n-1}(x',y')$ as:

$$\hat{I}_n(x,y) = \tilde{I}_{n-1}(f(x,y), g(x,y))$$

where $x' = f(x,y)$ and $y' = g(x,y)$ are the transformation functions.

The conventional approach of performing BMMC, relies on the assumption that the motion between the two frames is of a translational nature, i.e. the transformation functions for the pixels in the i_{th} block of the image are:

$$f(x,y) = x - \Delta x \text{ and } g(x,y) = y - \Delta y,$$

where $\mathbf{V}=(\Delta x, \Delta y)$ is the estimated translation vector for the i_{th} block. More sophisticated BMMC techniques, capable of handling motion other than only translation, can be composed by adopting different transformation functions. A broad class of such functions is the "*perspective transformations*" [8][5] (*affine* and *bilinear* are two other broad classes of such functions; however, we did not consider them in our research, because our own experiments have shown superior performance of the perspective transformations, see also [8]). The transformation functions for the pixels included in the i_{th} block of the image are:

$$f(x,y) = \frac{a_{i1}x + a_{i2}y + a_{i3}}{a_{i7}x + a_{i8}y + 1} \text{ and } g(x,y) = \frac{a_{i4}x + a_{i5}y + a_{i6}}{a_{i7}x + a_{i8}y + 1}$$

where $a_{i1},\ldots,a_{i8}$ denote the parameters of the transformation.

A significant improvement in the accuracy of the prediction for the new frame and thereby higher CRs can be achieved by using the perspective nonlinear mapping functions, especially if scenes undergoing complex movement, such as rotation, zooming, etc., are present in the frame. However, replacing the conventional translation only functions by the perspective ones is impractical, because of the increased computational complexity. On the other hand the simplicity of the conventional motion estimation algorithm and the superior performance of the perspective transformations can be combined in a hybrid scheme. Such a scheme, would be capable of handling complex movements while the complexity of the algorithm is still affordable. The hybrid scheme we used in our research works as follows:

- Each block i in the current frame is estimated using the conventional full search with half pixel accuracy block-matching algorithm.
- If the estimation error E_{CON} for the block i is less than or equal to T_{MAE}, where T_{MAE} is a sequence dependent threshold, the estimation is considered adequate and no further processing is required.
- If the estimation error is higher than T_{MSE}, then the block is also estimated using the perspective transformation and the estimation error E_{PER} is calculated. If E_{CON} - E_{PER} is greater than T_{TOL}, where T_{TOL} is a threshold used to favour the conventional BMMC algorithm, the current block is estimated using the perspective transformation functions.

In order to reduce the perspective computational complexity further, the classical three step search method [7] has been used to search the four corners of the current block, as explained in [8]. This strategy reduces, for a ±15 pixels/frame search area, the number of search quadrilaterals from 31^8 to 18849. Additionally, instead of coding the eight mapping parameters for each perspective matched block, the search index is coded. This is done using $ceil(log_2 18849)$=15 bits [8].

2.2 Controlling the video bit rate

The motion field is used to compute the DFD. This DFD is afterwards segmented into a number of regions using the segmentation algorithm described in [3]. One single segmentation is produced that is consequently used for the coding of all the components of the colour DFD image. The DFD is also used to produce a Binary frame, indicating the high and low energy regions [2][4]. Let N_{ki}, k=Y,U,V, be the number of pixels of the region i in the k channel of the DFD, n_{ki} the number of high energy points under the same region in the k channel of the Binary frame and T_k a fixed threshold for the channel k. The regions with

$$n_{ki} / N_{ki} \geq T_k, \qquad (1)$$

for at least one k=Y,U,V, are considered as moving regions and have to be coded, while the rest are considered as static regions and can be ignored.

The way regions are separated into moving and static can be used as a means to control the video bit rate. From eq. (1), it can be seen that if we set $T_k=0$, for all k=Y,U,V, all regions will be classified as moving and all of them will finally be reconstructed. This will give a better quality, but a much higher bit rate. On the other extreme, if we set $T_k>1$, for all k=Y,U,V, all the regions will be classified as static and none of them will be reconstructed. The bit rate drops in this case (only the motion vectors have to be transmitted), but the quality will be very bad, especially if large movements exist between coded frames.

Additionally, regions having larger n_{ki} / N_{ki} ratio values are more important than those with smaller ratio values, and should be reconstructed with higher priority. Therefore, based on their n_{ki} / N_{ki} ratio values, the moving regions can be ranked in terms of visual importance. The moving regions can then be coded in a hierarchical fashion starting from the most important one with the highest n_{ki} / N_{ki} ratio value. At each level of the hierarchy, the bits needed to code the moving regions, up to that level, are monitored and the total amount of bits to code the frame is estimated. Coding continues until all the moving regions have been coded or the available bandwidth for this frame has been reached. This scheme supports rate control, while it ensures that for a given bandwidth the most visually important regions will be reconstructed.

3. EXPERIMENTAL RESULTS

The well known *Miss America*, *Carphone* and *Foreman* sequences in QCIF format (176x144, 4:2:0 format, 25 frames/s) have been selected to illustrate the performance of the proposed techniques. In the experiments, the search range for the motion estimation was fixed at ±15 pixels. The compression of the DFD was done by using the weakly separable bases described in [9] and setting the maximum number of basis functions for each region to 16 and 4 for the luminance and chrominance channels respectively. The contours were simplified by the filter described in [1][4] before coding. In all cases 100 frames were used. Only one out of every three frames was coded (i.e. frame skipping of 2 frames), resulting in a decoded frame rate of 8.3 Hz.

Figures 1-3 depict the plot of the PSNR figures for the luminance channels, when the bit rate control strategy is used, for a transmission channel's bandwidth of 24 Kbit/sec, target frame rates of 10 frames per second (fps), 7fps and 4fps respectively, and using solely the conventional motion estimation technique. The results show a fair improvement of the PSNR with decreasing frame rate (results for *Miss America* are not shown, since the quality even at 10fps is very good for this sequence and no significant quality improvement is obtained with decreasing frame rate). The same conclusion can be drawn by looking at figure 7, where the coded frame 100 of the *Carphone* sequence for these three different bit rates is shown.

Figures 4-6 depict the improvement on the subjective image quality (in terms of PSNR) obtained by using the hybrid motion estimation (for a 24 Kbit/sec transmission bandwidth, 4fps) instead of the conventional block-matching. An average improvement of more than 1dB is observed in all the cases. Experiments have been also carried out, when no rate control is imposed. Similar improvement on the PSNR is demonstrated, while the bit rate is reduced by more than 10%.

4. CONCLUSIONS AND FURTHER RESEARCH

We have combined the simplicity of the conventional block matching motion estimation technique with the advanced prediction capabilities of the perspective transformation functions into a hybrid scheme capable to handle complex movements without excessive computational complexity. We have shown that the hybrid motion estimation scheme outperforms the purely conventional motion estimation method by improving the overall PSNR by more than 1 dB for reduced bit rate. A simple rate control strategy has been presented, in order to demonstrate that the way the regions are classified into "moving" and "static" can be used as a means to control the bit rate, ensuring at the same time that the most visually important regions will be reconstructed, for a certain bandwidth.

The inclusion of other parameters in the rate control strategy is currently under research. This will definitely allow a better regulation of the bit rate and the quality degradation. Ways to efficiently code the contour and the texture information of each region in the DFD, which will result in an embedded bitstream, are also investigated. This will allow the decoder to cut the bitstream at any point and therefore reconstruct the image at a lower bit rate, resulting in a progressive transmission video scheme.

5. ACKNOWLEDGMENTS

This work was supported by the IWT, under the "IT-TeleVisie" project ITA/950202/KUL.

6. REFERENCES

[1] V. A. Christopoulos, C. A. Christopoulos, J. Cornelis, and A. N. Skodras, "A New Contour Simplification Filter for Region-Based Coding", *Proceeding of the VIII European Signal Processing Conference (EUSIPCO-96)*, Vol. 1, Trieste, Italy, 10-13 September 1996, pp. 336-339.

[2] V.A. Christopoulos, C. A. Christopoulos, W. Philips, and J. Cornelis, "Segmented Image Coding with Contour Simplification for Video Sequences", *Proceedings of the International Conference of Image Processing (ICIP-96)*, Vol. 1, Lauzanne, Switzerland, 13-16 September 1996, pp. 693-696.

[3] V. A. Christopoulos, P. De Muynck, and J. Cornelis, "Colour Image Segmentation for Low Bit Rate Segmented Image Coding", *Proceedings of the 13th International Conference on Digital Signal Processing (DSP-97)*, Vol. 2, Santorini, Hellas, 2-4 July 1997, pp. 861-864.

[4] V. A. Christopoulos, P. De Muynck, and J. Cornelis, "Contour Simplification for Segmented Still Image and Video Coding: Algorithms and Experimental Results", *Signal Processing: Image Communication* (In Review).

[5] P. Heckbert, "Fundamentals of Texture Mapping and Image Warping", *Master's Thesis*, Dept. of EECS, University of California at Berkeley, June 1989.

[6] M. Kunt, M. Benard, and R. Leonardi, "Recent results in high-compression image coding", *IEEE Trans. Circuits and Systems*, Vol. 34, November 1987, pp. 1306-1336.

[7] H. G. Musmann, P. Pirsch, and H. J. Grallert, "Advances in Picture Coding", *Proc. of the IEEE*, Vol. 73, No. 4, April 1985, pp. 523-548.

[8] V. Seferidis, and M. Ghanbari, "General approach to block-matching motion estimation", *Optical Engineering*, Vol. 32, No. 7, July 1993, pp. 1464-1474.

[9] W. Philips, "Fast coding of arbitrarily shaped image segments using weakly separable bases", *Optical Engineering*, Vol. 35, Jan. 1996, pp. 177-186.

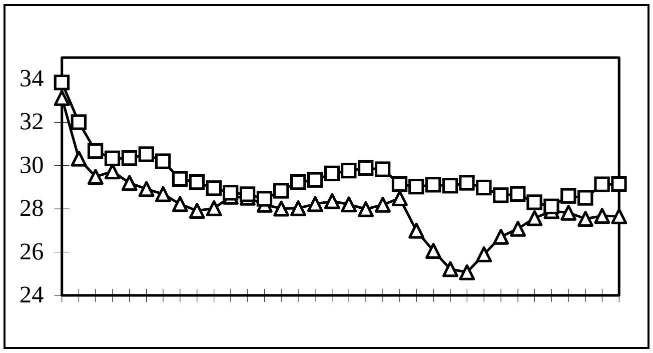

Figure 1. PSNR figures for a target frame rate of 10fps on a 24Kbit/s transmission channel (–– Carphone, –Δ– Foreman).

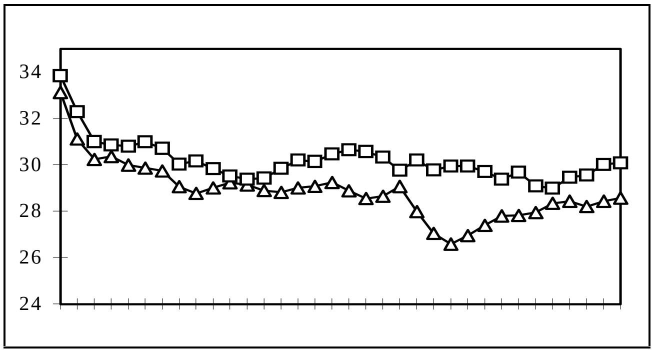

Figure 2. PSNR figures for a target frame rate of 7fps on a 24Kbit/s transmission channel (–– Carphone, –Δ– Foreman).

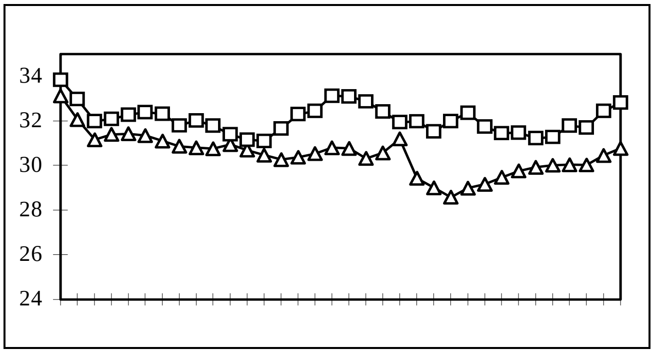

Figure 3. PSNR figures for a target frame rate of 4fps on a 24Kbit/s transmission channel (–– Carphone, –Δ– Foreman).

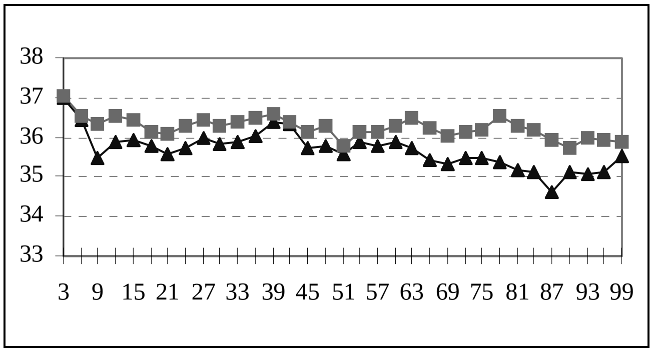

Figure 4. Hybrid vs. conventional motion estimation in terms of PSNR for the *Miss America* sequence (–– Hybrid, –Δ– Conventional).

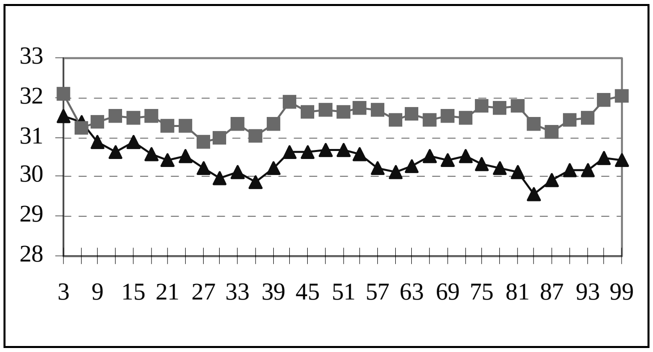

Figure 5. Hybrid vs. conventional motion estimation in terms of PSNR for the *Carphone* sequence (–– Hybrid, –Δ– Conventional).

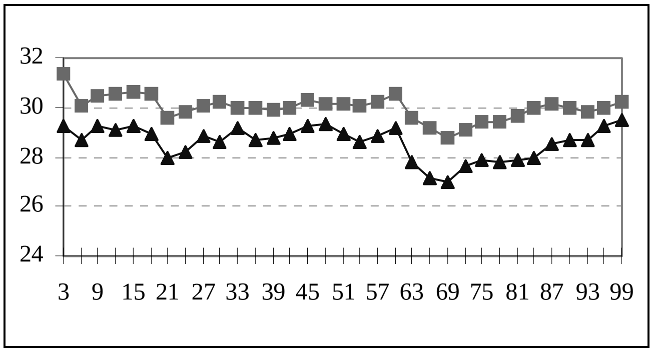

Figure 6. Hybrid vs. conventional motion estimation in terms of PSNR for the *Foreman* sequence (–– Hybrid, –Δ– Conventional).

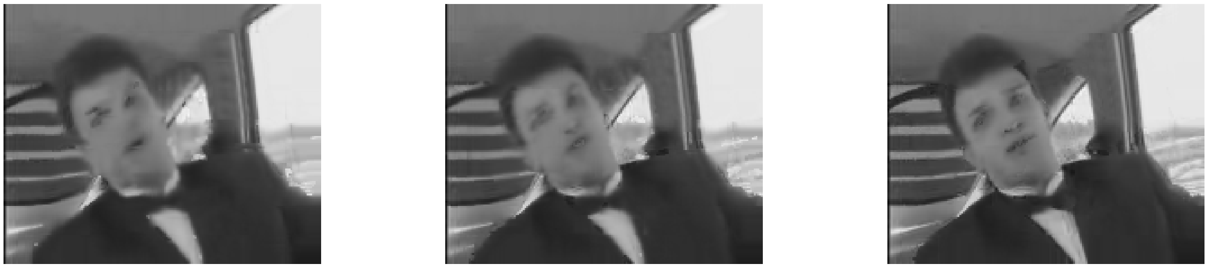

Figure 7. Coded frame 100 of the sequence *Carphone* (24Kbit/s, 10fps, 7fps, 4fps)