Multivector Motion Description for Region-based Very Low Bit-rate Video Coding

Luis Salgado, José I. Ronda, José M. Menéndez, Enrique Rendón, and Alberto Sanz
Grupo de Tratamiento de Imágenes, E.T.S. Ingenieros de Telecomunicación
Universidad Politécnica de Madrid, E-28040 Madrid, Spain
Contact e-mail: lsa@gti.ssr.upm.es, http://www.gti.ssr.upm.es

ABSTRACT

In the present paper, a new approach to region motion description and estimation is introduced, which results particularly suitable for segmentation-based coding strategies for very low bit-rate video coding. Region motion is described through a variable number of motion vectors (MV's) applied to specific control points. No information about this control points is required to be transmitted, as their determination is based on information available at the decoder. Results show an important net bit-rate saving for QCIF images using this new approach versus the standard translational model. Transmission at rates below 64 kbit/s with very high image quality are achieved.

1 INTRODUCTION

Motion description is a key point in most of the different approaches to region-based video coding for very low bit-rate transmission [3]. Among them, segmentation-based hybrid video codecs appear as an interesting alternative, where image analysis techniques are merged with a traditional video coding scheme in order to provide advanced functionalities (selective coding, user interaction, etc.) along with high coding efficiency. Typically, in these approaches information related to segmentation (region description), motion estimation, and prediction error has to be transmitted to the decoder.

In our system approach, region description is avoided through a combined forward/backward motion estimation strategy [1]. In this context, the coding of video at very low bit-rates can only be achieved with the help of advanced region-based prediction techniques which make unnecessary the transmission of the prediction error (DFD) excepting for very small areas of the picture [2].

In the present paper, a new approach to region motion description and estimation is introduced, which results particularly suitable for the abovementioned scheme.

2 REGION-BASED CODING WITHOUT CONTOUR TRANSMISSION

The region-based coding scheme without contour transmission, introduced in [1] (figure 1), operates by first

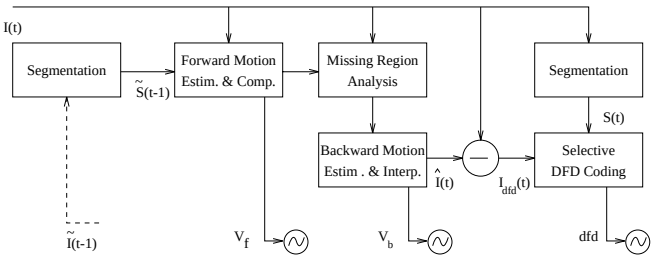

Figure 1: Block diagram of the region-based hybrid video codec.

segmenting a reference image $\tilde{I}(t-1)$ (the previous reconstructed image) and then applying a mixed forward/backward motion estimation strategy. Forward motion estimation is carried out by finding a correspondence for each region of the reference image $\tilde{S}(t-1)$ (also available at the decoder) in the image to be coded $I(t)$. For those areas in this image which are not covered by the mapping of the reference image regions, predictions are obtained by backward motion compensation. As the previously decoded image is used as the reference image, the segmentation can be performed by the decoder, making unnecessary the transmission of the region contours. Prediction error is segmented and selectively coded, providing the required flexibility to be adapted to very low bit-rate transmissions.

The advantages of removing the contour information from the video stream extend beyond the obvious saving in bit-rate, as the segmentation operation results also alleviated from the requirement of providing smooth, easy to code contours.

3 MULTIVECTOR MOTION DESCRIPTION

The limitations of the standard rigid translation model of the motion of a region are well known. On the other hand, the alternatives to this approach that can be found in the literature are oriented to the description of the evolution of large regions, and their application for small areas would result very expensive both in terms of computational cost and transmission bit-rate. The affine transformation [4], for example, requires the

transmission of six parameters, three times more than the rigid translation model.

Consequently, there is a clear need for region motion description methods which (a) are scalable in the number of parameters, depending on the size of the region and the complexity of its motion, (b) are able to cope with a wide variety of motion types and reshaping of the regions, and (c) allow for the computation of the parameters with a reasonable cost, taking into account the great number of regions in the same image to which the model has to be fitted.

Multivector motion description (MMD) is an alternative devised in order to meet these requirements which consists of two independent but perfectly integrating components: a syntax for the description of the motion of a region and, a low computational cost algorithm for the direct obtention of the motion parameters. The key of the approach consists in the adaptive selection for each region of one, two or more points, named control points, for which the motion is transmitted, and the obtention of the motion for the region through the adjustment of a suitable motion model to the transmitted data.

Depending on the number of MV's and the type of motion vector interpolation in the receiver, the approach can be made equivalent to different classical motion models or extend these models to include wider forms of temporal evolution of the region. More specifically, the following correspondence has been found suitable:

- One motion vector: rigid translation.
- Two motion vectors: isomorphic transform
- Three motion vectors: affine transform
- Four motion vectors: perspective projection

In our implementation each region is adaptively assigned a variable number of vectors on the basis of its size and contribution to the global prediction error. In the case of a single MV, the standard translational model results, and the vector is obtained by region matching. In the case of more than one MV, the control points are selected so that a compromise is achieved between the two following conditions:

1. Any pair of control points are far-apart enough from each other to show different MV values in case of non-translational motion.
2. The distance between any pixel of the region and the nearest control point is not excessive, so that the motion of each region point can be considered highly correlated with that of the nearest control point. Therefore, no large errors after motion interpolation are achieved.

Together with these two conditions, another restriction is imposed to keep minimum the increment of the amount of motion information due to the use of MMD: the algorithm to determine the control points must be based on information available at the decoder. Therefore, control points coordinates could be directly recovered at the decoder, so that no transmission of information apart from the MV values is required.

In the computation of the MV's, a locally translational model is applied [5]. Computation accuracy will rely on the control point neighbourhood considered for the estimation. Motion interpolation for the other pixels of the region is carried out by adjusting the estimated MV's to an affine model. Different restrictions are imposed to this model depending on the number of transmitted MV's. In this way, the expensive gradient-based search computation of the model parameters is avoided.

3.1 Control points determination

As stated before, our system approach to region-based video coding for very low bit-rate transmission avoids the transmission of any region description through a combined forward/backward motion estimation strategy. Motion estimation and compensation is based on regions, whose determination is strictly based on information available at the decoder. The same philosophy is applied to the determination of the control points, and the algorithm is based mainly on the regions shape.

Stemming from the two conditions stated in the previous section to drive the control points selection, these control points can be understood as representative points of the motion of region pixels closer to them. Thus, given a region R, the determination of N control points will lead to a partition where a region pixel is assigned to the closest control point.

In order to determine this partition, a region-growing strategy is applied. The starting points for the process (initial control points) are fixed as the N region contour points whose inter-distance sum is maximum (points are far enough to likely show different motion in case of non-translational motion). Region pixels are then iteratively labeled as belonging to one of the control point subregions keeping always two conditions: connectivity to pixels already assigned to the subregion and, minimum distance to the control point. Control points are updated at each iteration, becoming the centroids of the labeled pixels.

A hierarchical block-based implementation of this general strategy has been carried out. Location of control points, as basically region-shape dependent, is performed on very low resolution segmented images. Further refinement on higher resolution images benefits of operating on restricted region areas. On the other hand, groups of connected pixels (blocks) are used instead of single pixels at each iteration, so the number of iterations is strongly reduced. Besides, parallel implementations can be easily undertaken to achieve real time operation.

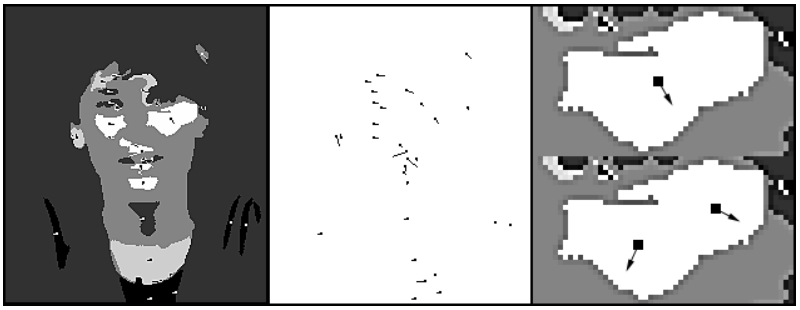

Figure 2: Image 1 of *Miss America*. *Left:* Segmentation with motion vectors on it. *Center:* Vector field using the standard translational model on the image regions. *Upper right:* Translational motion description of the white region under her left eye, showing the single MV. *Lower right:* Motion description of the same region showing the control points (square black blocks) and the two MV's.

4 RESULTS

Tests have been conducted on QCIF format *Miss America* (MA) and *Claire* (CL) sequences, with temporal resolution of 8 images per second. Analysis of the results will be focused on the application of the Multivector Motion Description (MMD) approach presented in this work to very low bit-rate video coding. In this context, the region-based video codec without contour transmission introduced in section 2 (figure 1) will be used.

The impact of using (MMD) instead of the simple translational model within the codec is studied in terms of: the quality of the resulting motion compensated images ($\hat{I}(t)$ in fig. 1), the amount of motion information to be transmitted (V_f and V_b in fig. 1), and the resulting transmission bit-rate (motion information plus *dfd* in fig. 1) keeping constant the average quality of the reconstructed images. Quality measures are provided evaluating the similarity between original images $I(t)$ and either motion compensated ($\hat{I}(t)$) or reconstructed ($\tilde{I}(t)$). For the similarity measure, the peak signal to noise ratio (PSNR) is used, defined as:

$$10log(\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} 255^2 / \sum_{n=0}^{N-1}\sum_{m=0}^{M-1}[p_t(n,m) - \bar{p}_t(n,m)]^2) \quad (1)$$

where N, M are respectively the number of rows and columns of the images, and the pixel difference between the original image (p_t) and either the compensated ($\bar{p}_t = \hat{p}_t$) or reconstructed ($\bar{p}_t = \tilde{p}_t$) images is considered as the error.

The average number of regions resulting from the segmentation are: 50 for *Miss America* and 60 for *Claire*. The use of the simple rigid translational motion model would imply a single MV per region for forward motion compensation (V_f). Thus, an average number of 50 and 60 motion vectors per image is used for forward motion compensation of MA and CL. Backward motion compensation is reduced to a few number of regions because most of the uncovered regions are motion interpolated.

The use of the MMD approach forces to decide to which regions control points and multiple MV's estimation will be carried out. Three types of information are evaluated to take this decision: the region size, the prediction error per region pixel achieved after motion compensation, and specially for very low bit-rates, a motion activity detector. For very small regions, the MMD will not be appropriate because the estimated motion vectors would be likely noise affected, while for very large regions, region partitions will be too large to assume motion correlation for some groups of region pixels. On the other hand, the activity detector allows to drive the application of the MMD to those regions which likely will better benefit from a more accurate motion compensation strategy.

Tests results present an average 20 % for MA and 25 % for CL of the regions whose motion is described through multiple MV's, implying around a 35 % increment in the number of MV's to be transmitted for both sequences. In terms of the number of bits per region devoted to motion information coding, the increase is not proportional to the increase in the number of transmitted MV's. The high correlation generally present among the MV's estimated to describe a region motion can be exploited to highly efficiently code them, for example, using advanced predictive MV coding strategies. The figure 2 shows the result of applying MMD to a sample region of *Miss America*.

Concerning the resulting motion compensated images ($\hat{I}(t)$), the quality when using only the simple translational model is compared with that obtained applying MMD. The gain using MMD is in the average of 2 dB for MA and 1.25 dB for CL, being this difference produced mainly by the higher complexity of the claire movements. Figure 3 presents an example of the accuracy achieved when using MMD. The attention is cen-

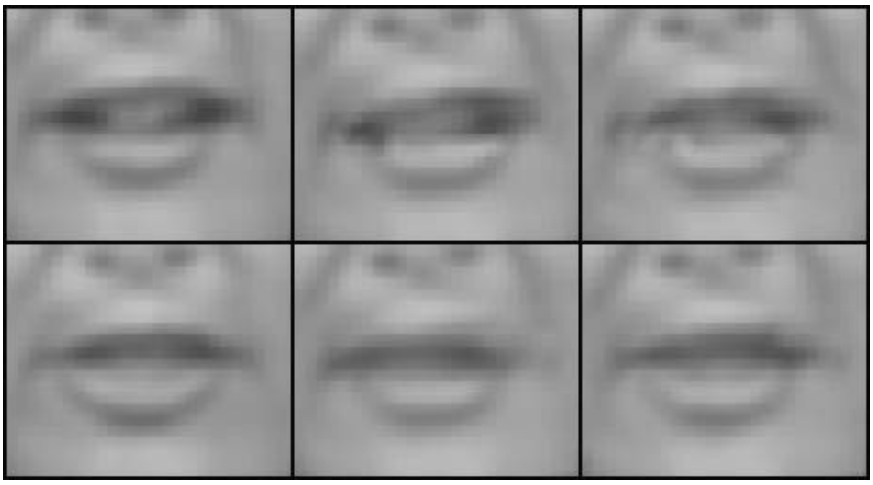

Figure 3: *Miss America* coded 8 imgs/sec at 48kb/s: mouth detail. *Upper-left:* Image 3 of the original sequence. *Lower-left:* Image 6 of the original sequence. *Center:* Motion compensated image 6. Movement is described with a single MV per region in the upper image, and with a variable number of MV's (between one and three) for the lower one. *Right:* Reconstructed image 6 at 48kb/s after motion compensation with one (Upper) or up to three (Lower) MV's.

tered on the mouth of Miss America in two consecutive images of the sequence to be coded. The left images show the original previous (upper) and current (lower) images to be coded. The center images present the result after motion compensation with a single MV description (upper) and applying MMD (lower). As it can be observed, with the simple translational model the mouth is almost not modified after motion compensation, looking very different from the target (lower-left). Applying MMD, the result after motion compensation (lower-center) is very similar to the target (lower-left). As a direct consequence, the prediction error with the translational model will be much higher than with the MMD (and so the number of bits required to code it). The right images of the same figure show the reconstructed images with one (upper) or up to three (lower) MV's.

Considering global transmission bit-rates, where bits devoted to motion information and prediction error coding are included, and a fixed average signal to noise ratio (PSNR) of 36 dB for MA and 35 dB for CL sequences, a net bit-rate saving above 15% for MA and 10% for CL has been found using the new motion description approach in comparison with the simple translational model. The small increment in the number of bits devoted to code motion information is clearly compensated by the final reduction in the transmitted prediction error.

5 CONCLUSIONS

A new approach to region motion description and estimation based on a variable number of vectors applied to specific control points has been presented. Control points transmission is dramatically eliminated, making this new strategy particularly suitable for video coding for very low bit-rate transmission. Results showing its applicability in the context of a segmentation-based hybrid video codec have been presented.

References

[1] L. Salgado, J. M. Menendez, N. García, E. Rendón, A. Sanz, "Segmentation-based Hybrid Video Codec for Very Low Bit-rates", Proc. WIASIC'94, 4 pages, Berlin, October 1994.

[2] Y. Yokoyama, Y. Miyamoto, M. Ohta, "Very Low Bit-rate Video Coding with Object-based Motion Compensation and Orthogonal Transform", Proc. Visual Communications and Image Processing '93, vol. 2094, pp. 12–23, Cambridge-Massachusetts, November 1993.

[3] L. Torres, M. Kunt, "Video coding: The Second Generation Approach", Kluwer Academic, 1996.

[4] Y. Nakaya, H. Harashima, "Motion Compensation Based on Spatial Transformation", Trans. on Circ. and Sys. for Video Technology, vol. 4 no. 3, pp. 339–356, June 1994.

[5] G. J. Sullivan & R. L. Baker, "Motion compensation for video compression using video grid interpolation", Proc. of ICASSP 91, pp. 2713–2716, Toronto, May. 1991.