SCALABLE SEGMENTATION-BASED CODING OF VIDEO SEQUENCES ADDRESSING CONTENT-BASED FUNCTIONALITIES

Josep Ramon Morros and Ferran Marqués

Dept. of Signal Theory and Communications
ETSETB - Universitat Politècnica de Catalunya
Campus Nord - Módul D5
C/ Gran Capità, 08034 Barcelona, SPAIN
Tel: (343) 401 74 04, Fax: (343) 401 64 47
E-mail: morros@gps.tsc.upc.es

ABSTRACT

In this paper, we address video scalability in the framework of a region-based coding system, allowing content-based functionalities. The proposed algorithm can construct scaled layers from a video sequence, each one with either fixed bit-rate or fixed quality, allowing content based manipulation. Two modes of operation have been defined: a *supervised mode*, that allows the user to select the objects to be coded in the enhacement layer, and an *unsupervised mode*, where this selection is done by the algoritm itself.

1. INTRODUCTION

Video coding scalability involves generating two or more video layers from a single video source. One of the layers, called basic layer, is encoded by itself to provide a basic representation of the image, while the other layers, when added progressively to the basic layer, produce increasing quality of the reconstructed signal. This functionality is useful for applications where the receiver display is either not capable or not willing to display the full resolution supported by all the layers.

Currently, many video coding standards support scalability. For example, in MPEG-2 there are four scalable modes, based on coding each layer with different sample rates or different spatial, frequential or temporal resolutions. All these techniques have in common the fact of dealing with the image at low level (pixel or block based). New video standards allowing content-based functionalities, as MPEG-4, will put other requeriments on scalability. In these coding schemes, the use of semantic-based objects, also named Video Objects, require new scalability techniques that can operate with individual objects of arbitrary shape in a natural way, maintaining as much as possible the coding efficiency. These techniques must be able to structure scalability layers on an object selection basis, rather that on pixel or block basis.

In this paper, we address video scalability in the framework of a region-based coding system [1, 9, 3]. This coding system is capable to describe the scene in terms of regions that can be coded independently. The proposed algorithm can construct scaled layers from a video sequence, each one with either fixed bit-rate or fixed quality. These layers are composed of regions at different resolution levels. The basic layer consists in a basic representation of the whole image, while the enhacement layers will code selected regions or objects as separate entities, improving its coded quality, and also allowing content based manipulation. Selection of the set of regions that will be coded at a given layer can be done by the user or by the algorithm itself. Note that regions and objects are not equivalent concepts. A region is an entity without semantic meaning that is formed for coding purpouses according to an homogeneity criterion. A Video Object is an entity with a semantic meaning. In our coder, a Video Object can have an internal structure and be composed of several regions. This internal structure results from the region-based nature of the coding algorithm. By ensuring that regions are formed respecting the contours of the existing Video Objects, objects are treated as regions or agrupations of regions. Therefore, content-based functionalities can be addressed in a natural way.

2. SCALABILITY IN A REGION BASED CODING SCHEME

The purpose of this section is to depict the problem of scaling a coded video source into several layers, such that each layer has associated a fixed bit-rate or fixed quality. Our goal is to provide scalability while mantaining the ability of the coding system to address content-based functionalities, such as independent retrieving and manipulation of objects. So, we want to construct scalability layers from a video source, maximizing its quality for a given coding rate. These layers should contain regions or objects present in the scene. A hierarchy of objects must be defined, to select which objects should be coded in the different layers.

In the case of video coding without scalability, only one coded video layer is generated. The generation of this single video layer can be faced as an optimization problem. In this case, the coding distortion must be minimized, given a restriction on the coding budget for a given set of quantizer levels. This problem has been widely adressed in source coding literature, both for block-based [10, 7] and for segmentation-based [4, 8] coding schemes. In this paper, we extend the coding scheme showed in [4, 6] to provide scalability. This extension can be done in two ways. The first is to define and code separately each layer. The second is to construct the layers in such a way that they are globally optimal in a rate distortion sense. While the second solution is desirable, it poses a dependent optimization problem, whose solution leads in general to exponential complexity algorithms. In

0-8186-8183-7/97 $10.00 © 1997 IEEE

this work, we have used the first approach.

The algorithm will provide in the basic layer a basic representation of the image. For the construction of the enhacement layers, two modes can be defined: *supervised region selection* or *unsupervised region selection*. In the unsupervised region selection mode, the optimization algorithm selects the appropriate regions that will be sent in each layer. In this work, the criteron to choose the regions is to optimize the coding of the layers for a given bit budget in a rate-distortion sense.

In the supervised region selection mode, Video Objects to appear in a given layer must be previously defined. In our tests for this mode, mask images have been used to define the objects to be included in each layer. These masks mark with different labels the objects to be encoded in each layer. Then, the regions that form these objects are tracked along the sequence and feed to the decision algorithm for appropriate selection [2].

3. SCALABLE CODING ALGORITHM

In this section we give a complete description of the scalability process. First, we outline the main steps of the algorithm. Then, a detailed description of each step is provided. As we treat scalability as an optimization problem, we will discuss briefly how this optimization approach has been used for video coding. In our scheme, scalability and coding are tighly related, and thus are described together. The overall coding algorithm is defined by three main steps:

1. **Projection step**. Partition of the current frame into regions and definition of meaningful objects. These regions and objects are related with the ones in already coded frames.
2. **Construction of the basic layer**. In this step a basic representation of the whole image is generated.
3. **Construction of enhaced layer(s)**. This is done by selecting the regions or objects in the scene that are to be coded in the enhacement layer(s).

3.1. Projection

We are interested in mantaining the temporal coherence of the regions along the successive frames. This is useful for coding purposes (motion compensation of contours and texture) and to address content-based functionalities. This can be done by deriving in the first instance the partition of the current frame from the partition of the last coded frame. To do this, motion information is applied to the partition of the previously coded frame #(n-1), to obtain a set of projected markers for the current frame #(n). These markers will be grown to achieve the final projected partition [5]. To appropriately address content-based functionalities, contours of any Video Object present in the scene must be imposed as a constraint in this process. This partition provides the time evolution of the regions in the previous partition, i.e., the tracking of the regions over each frame of the sequence. In this process some regions may disappear while new regions cannot appear. Segmentation for the current image is based on this partition.

3.2. Construction of the basic layer

In this step, a set of partition proposals is built, taking the projected partition constructed in the previous step as a basis. Some of the partitions in this set are created by merging regions of the projected partition on a motion similarity basis, while some others are built by re-segmenting this projected partition on a spatial basis [6, 4]. The result is a hierarchy of partitions, called the Partition Tree, that represents the image with various levels of detail. Variations with respect to the projected partition can be properly introduced by constructing the final partition for the current frame with a selection of regions from the different levels of the Partition Tree. All the regions of the tree are coded with a set of texture coding techniques, each one with various quantization levels. The resulting rate and distortion figures for each region are stored in a hierarchic structure, called the Decision Tree. An algorithm based on rate-distortion optimization using Lagrange multipliers [4] is used to select the regions that will form the basic layer partition. This algorithm also selects the optimal texture coding technique and quantizer choice for each region. Figure 1 depicts the whole process.

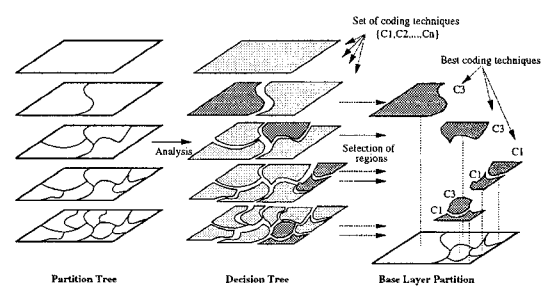

Figure 1: Decision process

3.3. Construction of the enhacement layers

The enhacements layers will improve the coding of the basic layer, while preserving the ability of the coding algorithm to address content-based functionalities. They are built by further refining the regions or objects already present in the basic layer or by defining new regions. As information about the texture of the regions has been already coded in the basic layer, the coding residue between the original image and the previous coded layer is used as information to be coded at the enhacement layers.

The construction of the enhacement layer takes as initial point the basic layer. The process is to build a new Decision-Tree taking the branches under the nodes selected in the base layer Decision-Tree. This new tree defines a modified Partition Tree (See Figure 2). That is, a new set of partition proposals that will be used by the algoritm to derive the final partition for the current layer. The same rate-distortion optimization process used to define the base layer is used here to build the enhacement layer. This step can be iterated to construct the desired number of enhacement layers.

At each layer, new regions or Video Objects will appear. The fact that the regions in the various levels of the Partition Tree are organized hierarchicaly is used by the optimization algorithm to select the appropriate regions or objects that

will be placed at the different scalability layers. This selection can be done by the coding algorithm itself, in order to obtain optimal results for a given criterion (in our case, rate-distortion). Another possibility is to allow interactivity of the user, by externally selecting the objects to be placed in each layer. In this case, the system is able to address content based functionalities.

- **Unsupervised region selection mode.** The objective is to enhace the coding already done in the basic layer by distributing the bit budget among the regions of the new Partition Tree. In this case, it is the optimization algorithm itself who decides which regions are to be sent in a given layer on a rate-distortion sense basis. The process is the same that is used for the basic layer. We have to point out that, in this case, regions may not have any semantic meaning, because the criterion that is used for its definition is purely coding efficiency. An example of this mode is shown in Figure 2.

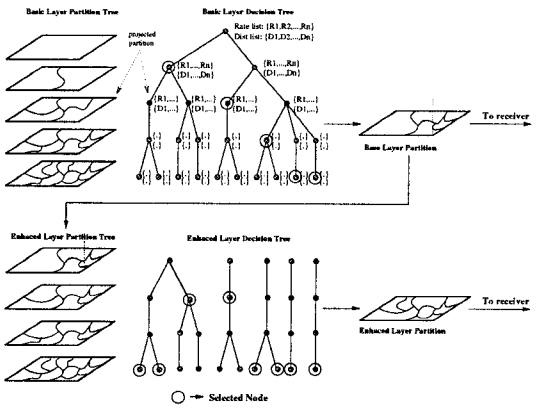

Figure 2: Unsupervised mode: Example with two layers

- **Supervised region selection mode.** In this mode, the key point is the possibility to provide external control or interactivity over the objects that are to be placed on the enhacement layers. To select the objects, a mask that marks all these objects and defines its hierarchy along the layers is used for each frame of the sequence. These masks can be externally provided for all the frames or, given a selection mask for the first frame of the sequence, the regions under the mask can be tracked along the sequence by means of an appropriate algorithm[2], so that the masks in the following frames are automatically generated. It is this capability of marking and tracking specific objects what gives the algorithm its strength to handle content-based functionalities.

4. RESULTS

In Table 1, the results of coding the sequences *News*, *Weather* and *Akiyo* with different bit rates are presented. The second column of the table shows the bit-rate devoted to encode the base layer. Third column shows the bit-rate for the enhacement layer and the fourth column shows the bitrate that takes the sequence at full resolution. This bit-rate is the sum of the bit-rates for the different layers. Finally, to give an idea of the quality, columns four and five show the PSNR of the selected object in the base and enhacement layers. This PSNR figure is an average over all the coded frames.

Sequ.	Base kbps	Enh. kbps	Full kbps	PSNR (base)	PSNR (enh.)
News	32.8	46.1	78.9	28.4	31.3
Weather	37.0	37.0	74.0	27.4	31.4
Akiyo	18.0	27.2	45.2	30.4	33.2

Table 1: Bit rates for three examples

Figure 3 depicts the quality of the coded frames along the sequence. For the *News* sequence, PSNR figures of the selected object in the base and enhaced layers are presented for each coded frame. Using 37.6 kbps for the enhaced layer, a quality improvement of 4.5 dB in average is reached.

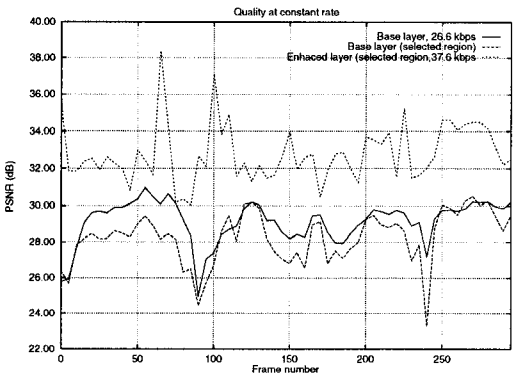

Figure 3: Evolution of PSNR of the news sequence.

Some results are presented using the supervised mode with two scalability layers. A mask image is provided for each frame of the sequence to define the objects included in the enhacement layer. Figure 4 shows, for a frame of sequence *News*, the original image, the base layer and enhaced layer coded images, as well as an example of the selection mask and resulting partitions for the two layers. Note that, in the enhaced layer, all the coding budget is spent on the regions inside the selection mask. It can be observed that the partitions for the base and the enhaced layers are quite different. The segmentation of the base layer reflect moving regions (e.g. the ballerina at the back screen), whereas the segmentation of the enhacement layer is constrained by the object selection mask. Note that the objects in the enhaced layer mask are composed of several regions. This re-segmentation is selected by the algorithm in order to optimize the coding of the objects.

5. REFERENCES

[1] M. Kunt, A. Ikonomopoulos, and M. Kocher. Second generation image coding techniques. *Proc. of IEEE*, 73(4):549–575, April 1985.

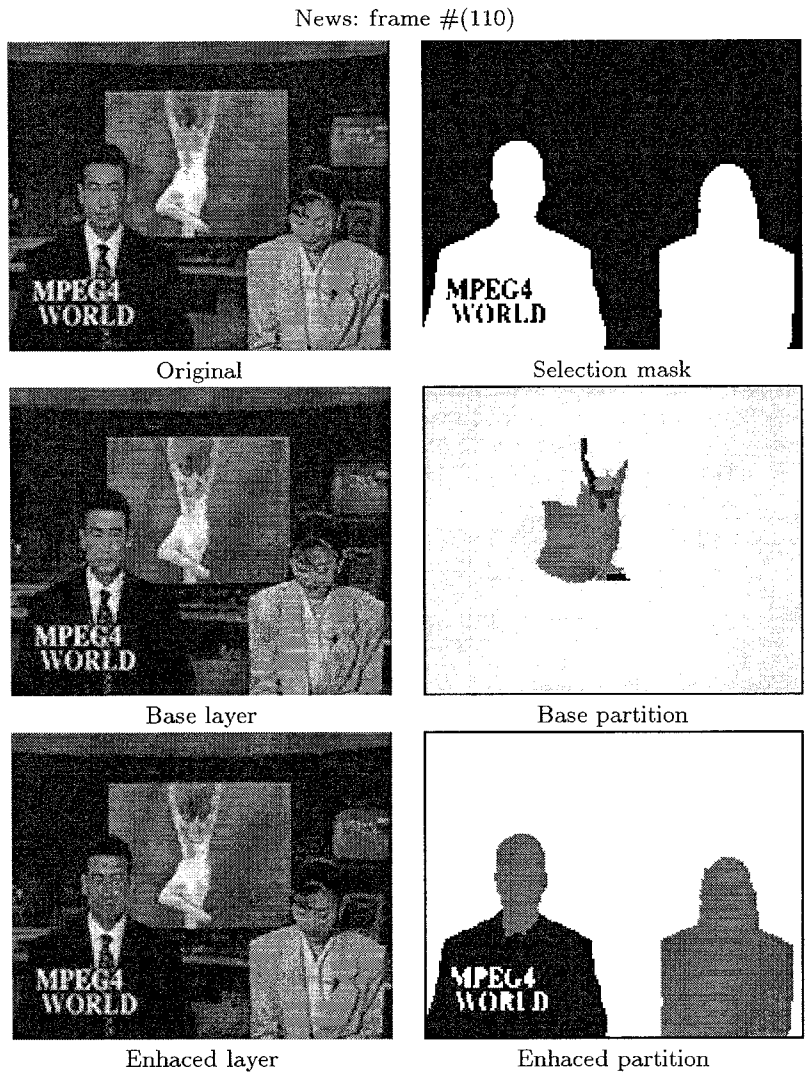

Figure 4: Coding examples

[2] F. Marqués and C. Molina. Object tracking for content-based functionalities. In *Visual Communications and Image Processing*, San Jose (CA), USA, Feb. 1997.

[3] F. Marqués, M. Pardàs, and P. Salembier. Coding-oriented segmentation of video sequences. In L. Torres and M. Kunt, editors, *Video Coding: The Second Generation Approach*, pages 79–124. Kluwer Academic Publishers, 1996. ISBN: 0 7923 9680 4.

[4] R. Morros, F. Marqués, M. Pardàs, and P. Salembier. Video sequence segmentation based on rate-distortion theory. In *SPIE Visual Communication and Image Processing, VCIP'96*, volume 2727, pages 1185–1196, Orlando (FL), USA, March 1996.

[5] M. Pardàs and P. Salembier. 3D morphological segmentation and motion estimation for image sequences. *EURASIP Signal Processing*, 38(2):31–43, Sept. 1994.

[6] M. Pardàs, P. Salembier, F. Marqués, and R. Morros. Partition tree for segmentation-based video coding. In *IEEE Int. Conf. on ASSP, ICASSP'96*, pages 1982–1985, Atlanta (GA), USA, May 1996.

[7] K. Ramchandran and M. Vetterli. Best wavelet packet bases in a rate-distorsion sense. *IEEE Trans. on Image Processing*, 2(2):160–175, April 1993.

[8] E. Reusens. Joint optimization of representation model and frame segmentation for generic video compression. *Signal Processing*, 46:105–117, September 1995.

[9] P. Salembier, L. Torres, F. Meyer, and C. Gu. Region-based video coding using mathematical morphology. *Proc. of IEEE*, 83(6):843–857, June 1995.

[10] Y. Shoham and A. Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Trans. on ASSP*, 36(9):1445–1453, Sept. 1988.