A PROPOSAL FOR DEPENDENT OPTIMIZATION IN SCALABLE REGION-BASED CODING SYSTEMS

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: (34) 934 016 550, Fax: (34) 934 016 447
E-mail: morros@gps.tsc.upc.es

ABSTRACT

We address in this paper the problem of optimal coding in the framework of region-based video coding systems, with a special stress on content-based functionalities. We present a coding system that can provide scaled layers (using PSNR or temporal content-based scalability) such that each one has an optimal partition with optimal bit allocation among the resulting regions. This coding system is based on a dependent optimization algorithm that can provide joint optimality for a group of layers or a group of frames.

1. INTRODUCTION

We present in this work an algorithm for optimal segmentation and bit allocation for dependent quantization in the framework of segmentation-based coding systems [1, 7, 2]. In previous works we dealt with the problem of optimal segmentation and bit allocation for independent quantization [3], and later we addressed content-based PSNR video scalability in a region-based coding system, while addressing content-based functionalities [4]. In those works, optimization was done independently for each frame or scalability layer; this is, without taking into account inter-frame or inter-layer dependencies. Now, we propose an algorithm to perform the segmentation and bit allocation optimization globally for a group of frames (in temporal scalability), or for a set of scalability layers (in PSNR scalability). We use an operational rate-distortion approach [5] to solve this dependent optimization problem. In addition, the system is extended by introducing content-based temporal scalability.

New video coding standards must provide a content-based representation of the visual data to allow the access and manipulation of the entities present in the scene. These entities, usually called Video Objects (VO), are parts of the image that have a semantic meaning by themselves. The fact that segmentation-based coding systems rely on a description of the scene in terms of regions leads to a natural representation of the Video Objects. Such VO's can be defined by selecting and tracking one or more regions in the scene whose contours match the contours of the VO. This can be done by allowing external interaction in the segmentation process.

For example, it is possible to define a Video Object in the first frame of the video sequence and to constrain the segmentation process to respect the contours of this object. The set of regions covered by the VO are marked and, due to the fact that the coding system preserves the temporal coherence of regions, the VO can be effectively tracked along the video sequence.

In many video coding applications, the encoding process should yield a bitstream which can be decoded by receivers with different display capabilities. 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, called enhancement 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.

The coding system we are presenting in this paper can support two different types of scalability: PSNR and temporal scalability. A complete description of our approach to PSNR scalability can be found in [4]. In temporal scalability two or more video layers are generated. In the basic layer, the sequence is coded at a low temporal resolution in order to provide a low bit-rate representation of the video sequence. In the enhancement layers, temporal resolution is improved by coding frames (or selected Video Objects in content-based scalability) that were skipped in the base

0-7803-5467-2/99/ $ 10.00 © 1999 IEEE

layer (see Figure 4 in Section 2).

This paper is organized as follows: In Section 2, a description of the encoding system is presented. Section 3 is devoted to formulate the dependent optimization problem and to present the algorithm that is used to tackle it. In Section 4 experimental results are presented and discussed. Finally, in Section 5, some conclusions are provided.

2. CODING SYSTEM

In this section we will briefly describe the region based coding system. A more detailed explanation can be found in [8, 4]. The coding system creates a description of the scene in terms of regions. To remove inter-frame redundancy, a prediction for both the partition and the texture of the current frame is created (*projection step*) using the last coded frame and motion information. In this projection process, the temporal coherence of the regions along the successive frames is preserved, so that they can be tracked through the video sequence. This projected partition is used to construct a set of partitions that describe the scene with various levels of detail. The motion, contour and texture information of these regions are then coded. Various texture coding techniques are used so the optimization algorithm can select the best technique and the best quantizer for each region, as well as the optimal partition for a given bit budget.

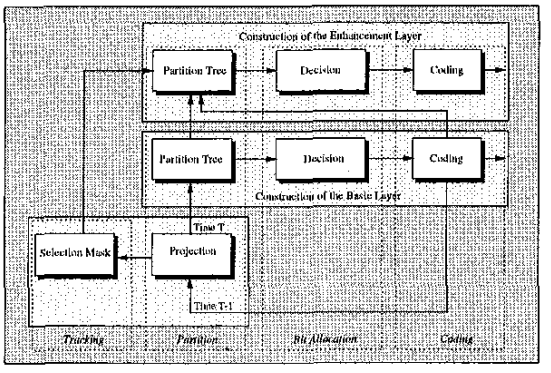

Figure 1: Encoding process

This coding system can support several types of scalability. In every case, each layer can be constructed so it has optimal quality for a given bit rate. For simplicity, we will consider only the case of two scalability layers: a basic and an enhancement layer. The extension to n-layers scalability is straightforward. Three main steps can be outlined (See Figure 1):

1. Projection of the basic layer. Partition of the current frame into regions and definition of meaningful objects. These regions and objects are related to those in the basic layer of already coded frames.

2. Construction and coding of the basic layer. Taking as a basis the projected partition constructed in the previous step, a hierarchy of partitions (*Partition Tree*) is constructed by splitting or merging regions. These partitions represent the scene with various levels of detail. All the regions in all the levels are coded. Various intra-frame and inter-frame coding techniques are used and in each case various quantization levels are available for each technique. The resulting rate and distortion figures are stored in a hierarchical structure (*Decision Tree*) that is used by the R-D optimization algorithm to find the optimal representation of the image, that is, to select regions from different levels to form the final partition, and the quantizer choice for each region.

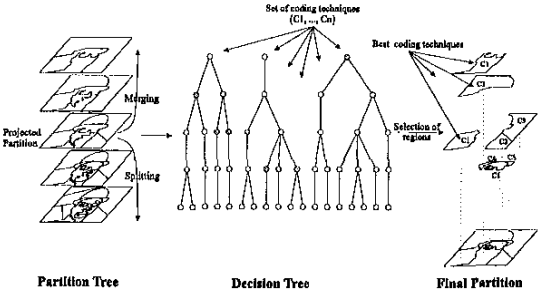

Figure 2: Decision process

3. Construction and coding of the enhancement layer. If more than one layer is to be generated, a process similar to the previous one is used. Nevertheless, for each type of scalability, specific adaptions are necessary.

In *PSNR scalability*, a new partition tree is constructed taking as a basis the basic layer coded partition and the basic layer partition tree. The goal is to improve the coding already done in the basic layer. Two possibilities are considered: The first one is to improve the coding of the texture of regions present in the basic layer partition (for example, by computing and coding the error between original and basic layer coded images). The second one is to refine the partition by introducing new regions. The new partition tree is constructed with the regions of the basic layer PT that represent the scene with equal or better resolution than the ones selected in the previous layer partition. This is, taking the branches under the selected nodes in the previous layer PT (see Figure 3). Then, the same optimization algorithm used in the basic layer is applied to find the final partition and quantizer choice for each region.

In *temporal scalability*, the basic and the enhancement layers represent different frames (see Figure 4). In this case, the enhancement layer partition tree is not a subset of the basic layer partition tree. Instead, the same process used to construct the basic layer is applied: Projection of the coded basic layer partition, construction of the partition tree and optimal coding of the texture and the partition.

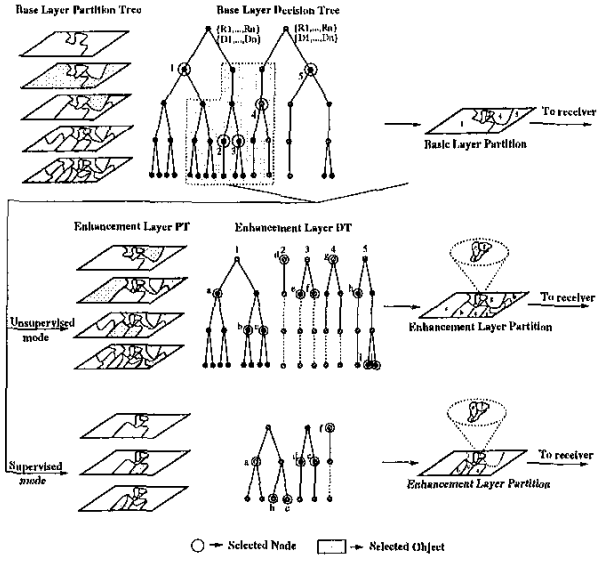

Figure 3: PSNR scalability: Construction of the new PT

Temporal object scalability has to deal with the possible apparition of uncovered background zones. This occurs when the shape of the object coded in the enhancement layer of frame #(n+k) is different from its reference in the basic layer of frame #(n) (see Figure 4). This is solved by detecting these uncovered background zones, that are then filled by motion compensating the texture of the neighboring regions. This way, only few motion vectors are used to code these non selected areas and almost all the available bit budget can be spent on the coding of the object itself.

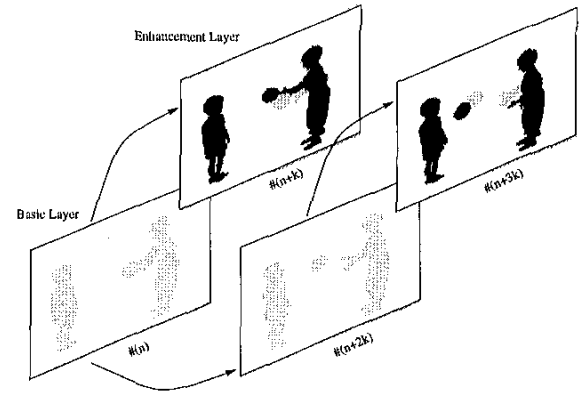

Figure 4: Temporal Scalability

In both types of scalability, it is possible to define Video Objects in a supervised or unsupervised way by giving a region or set of regions a semantic meaning. This allows to select the objects that are to be placed on the enhancement layers (*object scalability*). This way an object can be defined and tracked through the video sequence in the enhancement layer, ensuring that all the available bit budget is spent only on the refinement of this object. In Figure 3 we can see the differences between supervised and unsupervised modes. In unsupervised mode the optimization algorithm selects wich regions will form the final partition in order to improve the coding of the entire frame. In supervised mode, an VO is defined (head & shoulders in the example) and the construction of the PT is restricted to the regions covered by this VO. New regions can be selected inside the VO if this is optimal in a rate-distortion sense (see detail of the head).

3. DEPENDENT OPTIMIZATION

The coding system that is depicted in Section 2 has temporal and spatial dependencies. The temporal dependency arises from the fact that the partition for the current frame depends on the partition of the previous frame. In the same way, the quality of the reconstructed image depends on the quantizers used in the current and previous frames. The spatial dependencies can be found in the construction of the scalability layers, where both the partition and the quality of a given layer depend on the previous coded layers.

Lets consider the case with two coding units. These coding units can be, for example, a set composed by a basic and enhancement layers representing the same frame, or two frames in temporal scalability, or a full frame and a VO in another frame in content-based scalability.

Let $\mathcal{P}$ be the set of possible partitions that are to be studied for optimality, $\mathcal{Q}$ the set of available texture coding techniques. Let $P_1, P_2 \in \mathcal{P}$, $Q_1, Q_2 \in \mathcal{Q}$ be a choice of partition and of the set of quantizers for each coding unit respectively. Let $D_1(P_1, Q_1)$, $D_2(P_1, Q_1, P_2, Q_2)$ and $R_1(P_1, Q_1)$, $R_2(P_1, Q_1, P_2, Q_2)$ be the associated distortion and bit rate. The global optimization problem to solve can be stated as:

$$\min_{Q_1,Q_2,P_1,P_2} [D_1(P_1, Q_1) + D_2(P_1, Q_1, P_2, Q_2)] \quad (1)$$

where

$$R_1(P_1, Q_1) + R_2(P_1, Q_1, P_2, Q_2) \leq R_{budget} \quad (2)$$

This is a highly complex constrained optimization problem, with exponential increasing complexity. Using Lagrangian relaxation, it can be converted to an equivalent unconstrained problem.

$$\min_{P_1,Q_1,P_2,Q_2} [J_1(P_1, Q_1) + J_2(P_1, Q_1, P_2, Q_2)] \quad (3)$$

where J_1 and J_2 represent the Lagrangian cost that results from merging rate and distortion through the Lagrange mul-

tiplier $\lambda \geq 0$.

$$\begin{aligned} J_1(P_1, Q_1) &= D_1(P_1, Q_1) + \lambda R_1(P_1, Q_1), \\ J_2(P_1, Q_1, P_2, Q_2) &= D_2(P_1, Q_1, P_2, Q_2) + \\ & \quad \lambda R_2(P_1, Q_1, P_2, Q_2) \end{aligned} \tag{4}$$

In [6] a similar approach was used, for example, to solve the temporally dependent problem for an MPEG GOP. However, in our system this solution is not feasible because in addition to the selection of the quantizer level for each region, the optimal partition for the image has to be constructed. This implies computing every possible partition (from the set of regions present in the partition tree) for each scalability layer, and then coding all the resulting regions with the available set of quantizer techniques. The complexity of this approach grows exponentially with the number of frames, the number of regions and the number of quantizer choices.

Moreover, in the PSNR scalability problem, it would be useful to introduce an additional constraint to ensure that the quality of the basic layer reaches a minimum, leading to a new and more complex problem:

$$D_1 \geq Q_{min} \tag{5}$$

The global optimization algorithm is adapted from [6], where it was applied to the case of pyramid-based multiresolution coding. It is based on the fact that, within each frame or scalability layer, all blocs operate at a constant slope at optimality. For the first frame or layer, the solutions on the convex hull are examined by sweeping the value of the Lagrange multiplier λ from zero to infinity. For each value of λ, the Lagrange optimization process is used on the first frame or scalability layer to find a partition of the image, with the optimal quantizer choice for each region. If the resulting rate satisfies that $R_{1,\lambda_i} \leq R_{budget}$, an optimization process is performed on the second frame or scalability layer in order to minimize its distortion for the bit budget $R_{2,i} = R_{budget} - R_{1,i}$. At the end of the process, the values that minimize $D_1 + D_2$ are chosen. The algorithm can be stated as:

Step 1: Starting at $\lambda = 0$, find $R_{1,\lambda=0}$ and $D_{1,\lambda=0}$. If $R_{1,\lambda=0} \leq R_{budget}$, optimize the second frame or scalability layer for the rate $R_{2,0} = R_{budget} - R_{1,0}$ and store the resulting values of $R_{1,0}, D_{1,0}, R_{2,0}, D_{2,0}$. Otherwise, sweep λ until a feasible value is found.

Step 2: Repeat the previous step by sweeping λ towards infinity. until $D_{1,\lambda_j} \geq D_{1max}$. The monotonicity properties ensure that for larger values of λ, the resulting solutions will not satisfy the restriction stated by equation 5.

Step 3: Choose the solution that minimizes $D_1 + D_2$.

The computational complexity of this approach is still high, though feasible for groups of two frames, VO planes or scalability layers.

4. EXPERIMENTAL RESULTS

Here, some results are presented to show the difference between independent optimization and two-frame global or dependent optimization. We restrict the dependent problem to two frames or scalability layers because of the computational complexity which is exponential in the dependency tree. For PSNR scalability, the results show that the global coding is performing better than the separate optimization. Figure 5 shows the quality of coded frames for PSNR scalability. For the News sequence, coded at 5Hz, 48 kbps for the basic layer, and 30 kbps for the enhancement layer, the average PSNR gain is more than 1 dB.

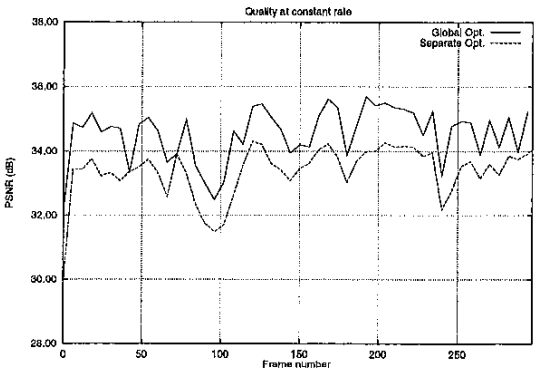

Figure 5: Evolution of PSNR of the News sequence.

In the temporal scalability case, the difference between dependent and independent optimization is much smaller. In Table 1, the average PSNR for all the frames of three sequences is depicted. We can see that little gain is obtained when using dependent optimization, at the expense of strongly increased computational complexity.

Sequence	Dependent Opt. (dB)	Independent Opt. (dB)
News (0.31 bpp)	33.19	33.04
Weather(0.42 bpp	33.02	32.91
Akiyo (0.33 bpp)	35.19	35.00

Table 1: PSNR in temporal scalabilty

Figure 6 shows the strategy adopted for uncovered areas in temporal object scalability. In the first row, the masks that define the selected video object in the base (a) and enhancement (b) layers and the object coded in the enhancement layer (c) are shown. To form the final coded image, the enhancement layer is added to the basic layer, where the selected object has been removed (d). The uncovered background is shown in (e) and the final result, in (f).

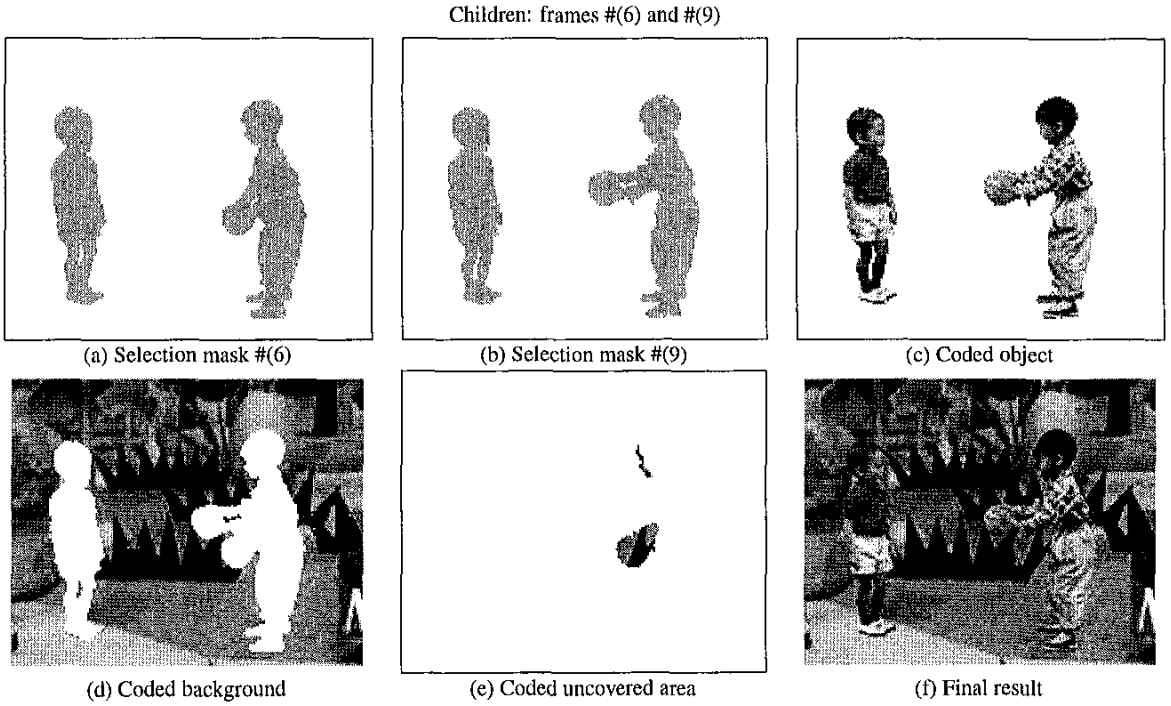

Figure 6: Object temporal scalability

5. CONCLUSIONS

We have shown an dependent optimization strategy to address the problem of bit allocation within a segmentation-based video coding system with PSNR and temporal scalability support. A big gain is obtained in PSNR scalability when comparing with an independent optimization algorithm. In temporal scalability, the improvement is marginal and may not worth the increase in complexity. As a future work, we plan to extend this coding system by adding spatial scalability, using a combination of PSNR and temporal scalability techniques.

In addition to optimal coding, one of the strengths of the video coding system is its ability to address content-based functionalities in a natural way due to its region-based nature.

6. REFERENCES

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

[2] 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.

[3] J.R. Morros, F. Marqués, M. Pardàs, and P. Salembier. Video sequence segmentation based on rate-distortion theory. In *SPIE VCIP'96* pp. 1185–1196

[4] J. R. Morros and F. Marqués. Scalable segmentation based coding of video sequences addressing content-based functionalities. In *IEEE ICIP'97*, vol II, pp. 1–4.

[5] A. Ortega and K. Ramchandran, "Image and video compression," *IEEE Signal Processing magazine*, vol. 15, no. 6, pp. 23–50, November 1998.

[6] K. Ramchandran, A. Ortega, and M. Vetterli. Bit allocation for dependent quantization with applications to multiresolution and mpeg video coders. *IEEE Trans. on IP*, 3(5):533–545 Sep. 1994.

[7] 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.

[8] P. Salembier et al. Segmentation-based video coding system allowing the manipulation of objects. *IEEE Trans. on CSVT*, pp 60-74 Feb.1997.