Segmentation of video sequences and rate control

Beatriz MARCOTEGUI *
Ferran MARQUÉS **
Joseph Ramon MORROS **
Montse PARDÀS **
Philippe SALEMBIER **

Abstract

This paper deals with the relation between segmentation for coding and rate control. The efficiency of a segmentation-based coding scheme heavily relies on this step that defines how many and which regions have to be segmented. In this paper, we show that this problem can be formulated as a rate/distortion problem. The proposed solution not only controls the segmentation, but also defines the coding strategy to be used in each region. Together with the general approach, several simplified versions of the segmentation control are proposed and discussed.

Key words : Image processing, Moving image, Segmentation, Image coding, Optimization, Transmission, Transmission rate, Decision tree.

SEGMENTATION DE SÉQUENCES VIDÉO ET MAÎTRISE DU DÉBIT BINAIRE

Résumé

Cet article traite de la relation entre segmentation pour le codage et maîtrise du débit binaire. L'efficacité d'un système de codage à base de segmentation repose largement sur cette étape qui définit le nombre de régions et quelles sont celles qui doivent être segmentées. Cet article montre que ce problème peut être formulé comme un problème d'optimisation de débit/distorsion. La solution proposée non seulement détermine la segmentation, mais définit également la stratégie de codage qui doit être utilisée dans chaque région. En parallèle avec l'approche générale, plusieurs approches simplifiées sont proposées et discutées.

Mots clés : Traitement image, Image animée, Segmentation, Codage image, Optimisation, Débit transmission, Arbre décision.

Contents

I. INTRODUCTION

In the two previous papers of this issue *Segmentation of video sequences for partition tree generation* [5] and *Bottom up segmentation of image sequences for coding* [2], various segmentation techniques have been described. They all involve two important features : first, they segment the video sequence in a coherent way, that is, they relate the regions of past frames to regions of the current frame (*projection* step). Second, they do not define a single segmentation but a set of regions that can be used in order to create the final

* Centre de Morphologie mathématique, Ecole des Mines de Paris, 35, rue Saint-Honoré, F-77305 Fontainebleau Cedex, France. Marcotegui@cmm.ensmp.fr.
** UPC, Campus Nord, Modulo D5, c. Gran Capita, s/n, 08034 Barcelona, Espagne. ferran@gps.tsc.upc.es, morros@gps.tsc.upc.es, montse@gps.tsc.upc.es, philippe@gps.tsc.upc.es.

partition. In the sequel, this set of regions is called the *universe of regions*. The universe of regions should include regions that are highly correlated with regions in previous frames, but also allow the introduction of new regions describing objects that were not present on the previous frames. Furthermore, some objects of the scene may disappear and their corresponding regions should be eliminated. The universe of regions can be viewed as a limited set of regions that may belong with high likelihood to the final partition. This paper focuses on the problem of finding the partition that best represents the image.

This issue of finding the best partition is highly linked to the rate control problem. Indeed, a very fine segmentation, involving a high number of regions, can in general well represent the picture but the resulting coding cost is very high. For coding applications, the problem is generally not to find the best segmentation but the best segmentation that leads to a coding cost below a given limit.

Furthermore, the definition of the regions should not only be based on the characteristics of their pixels (homogeneity in some sense) but also on the set of coding techniques that can be used to represent them. As a result, the encoder structure should not rely on the *traditional* two steps approach : analysis and *then* coding. A third step should be introduced to make the link between the analysis and the coding blocks. In the sequel, this third step will be called the *decision* (see Fig. 1).

In this paper, several approaches for defining the best coding-oriented partition out of the universe of regions are presented and discussed. The first one is based on the theory of rate/distortion optimization and is presented in the next section. Alternative techniques using simplified rules are presented in Section III.

II. RATE/DISTORTION OPTIMIZATION

II.1. Partition tree and decision.

In this section, we concentrate on the coding structure illustrated in Figure 1. As discussed in the introduction, the encoding process relies on three sets of functions : bit allocation function, partition functions and coding functions.

Bit allocation function : this function, corresponding to the *decision* block of Figure 1, makes the link between the analysis (partition) functions and the coding functions of the encoder. This processing step represents the main subject of this section. It defines the coding strategy, that is the set of regions to be coded a well as the type of coding technique to be used in each region. The coding strategy is obtained as a result of an optimization in the rate/distortion sense.

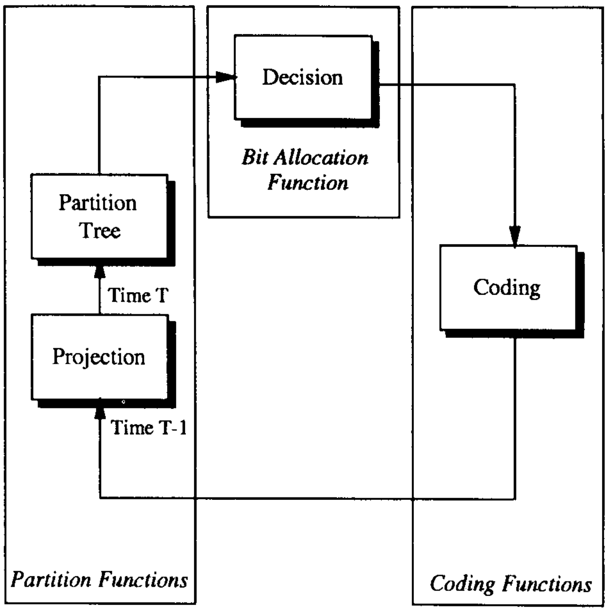

FIG. 1. — Scheme of the segmentation-based coder.

Diagramme du codeur à base de segmentation.

Partition functions : the objective of this set of functions is to create for each frame a universe of regions out of which the *decision* has to create the final partition. Note that, to be able to track objects, this universe of regions should be related to the previous partition. This is the objective of the *projection* block of Figure 1. Moreover, the universe of regions has to offer to the *decision* the possibility to introduce new regions or to eliminate past regions. This function is achieved by the *partition tree* block [6, 5, 2].

Coding functions : the last set of functions actually codes the information necessary to restore the sequence on the receiver side. It deals with the encoding of the coding strategy, the motion information, the partition and the pixel values.

Figure 2 summarizes the decision process : based on the rate/distortion criterion, the *decision* block selects the best strategy in terms of regions and coding techniques among a set of possibilities. The *partition tree* contains all regions that may belong to the final partition; for each of them, a set of possible coding techniques $\{C_1, \cdots, C_n\}$ is proposed to the *decision.* This last set involves several region-based coding techniques with various levels of quality. Moreover, the techniques can be proposed in intra-frame mode (coding of the original signal) and in inter-frame mode (motion compensation of the region and coding of the prediction error). Then, taking regions from various levels of the *partition tree*, the *decision* block defines jointly the best partition and the best coding technique for each region.

II.2. Rate/distortion optimization.

The definition of the best partition and the coding strategy can be formulated as a constrained optimization problem in the rate/distortion sense. Let us first review the problem of constrained bit allocation.

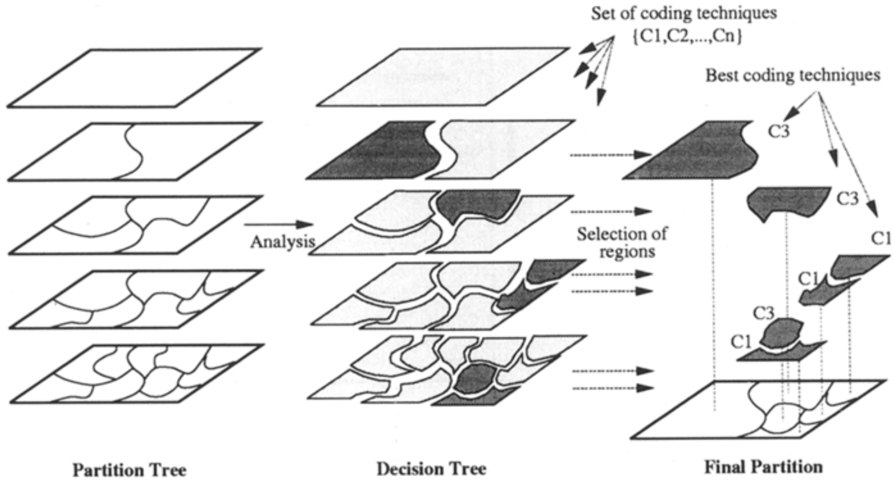

FIG. 2. — Decision process.

Processus de décision.

The problem of minimizing the distortion D subject to a restriction on the coding cost R has been widely addressed in source coding literature [11, 7, 8]. The constrained problem can be formulated as follows : given an image $\mathcal{I}$, a budget R_{budget}, and a set of coding techniques $Q = \{q_i\}$, find the best coding strategy $q^* \in Q$ such that :

$$(1) \quad D(q^*) = \min_{q_i \in Q} D(q_i) \text{ subject to } R(q_i) \leq R_{budget}.$$

This constrained problem can be converted into an unconstrained problem by combining the distortion and the rate by means of a Lagrange multiplier λ, and then, minimizing the Lagrangian cost function $J(\lambda) = D + \lambda R$. It can be shown [11] that, for each $\lambda \geq 0$, the solution $q^*(\lambda)$ of :

$$(2) \quad \min_{q_i \in Q}(J(q_i)) = \min_{q_i \in Q}(D(q_i) + \lambda R(q_i)) \quad \lambda \geq 0,$$

is the solution of the constrained problem with $R_{budget} = R(q^*(\lambda))$.

The problem is therefore to find the optimum Lagrange multiplier λ^* such that $R(q^*(\lambda^*)) = R_{budget}$ and the corresponding optimum coding technique $q^*(\lambda^*)$ defined by equation (2).

II.2.1. Finding the best coding strategy.

In this section, we assume that the Lagrange parameter is known and we discuss the definition of the best coding strategy. The optimization process relies on the concept of *decision tree* [8] illustrated on Figure 3. This tree concentrates in a compact and hierarchical structure all the possible coding choices. The *partition tree* defines the choices in terms of regions. The list of coding techniques deals with the actual coding of these regions.

The structure of the *decision tree* is defined by the *partition tree* [3] : each node of the *decision tree* corresponds to a region in the *partition tree*. The relations father/children between the nodes are also given by the *partition tree* and define how one region at a given level may either be split in various regions (children) or be merged to form a larger region (father). To define the coding strategy in the rate/distortion sense, the *decision tree* should also convey the information about the coding cost (rate measured in number of bits) and quality (distortion assessed by the squared error) of all possible coding techniques. Therefore, a list of rates (*rate list* in Fig. 3) and a list of distortion (*dist list* in Fig. 3) are assigned to each node. In practice, each region of the *partition tree* is coded by all techniques (with various quality levels and either in intra-frame mode or in inter-frame mode) and the corresponding rate and distortion values are stored in the *decision tree*. The computation of the distortion is rather straightforward. We have used the squared error between the original and coded frames (that is the sum of squared difference between the values of the original and coded frames for all pixel belonging to the region support). It is however not the case for the computation of the rate. Indeed the rate associated to a region is composed of the sum of the number of bits devoted to the texture, the motion as well as the shape information. In practice, the texture and the motion information is coded independently for each region so there is no major difficulty to define the texture or the motion rate for each region. But this independence is not maintained for the shape information because a contour is always shared by two regions. In order to avoid the problem of optimization with complex dependency between regions, we have used the following approximation of the shape rate : based on the analysis of a large number of coded frames, an average number of bit per contour points has been computed and the shape rate assigned to a region is equal to this average figure multiplied by the region perimeter divided by two (each contour point is shared by two regions).

Finally, note that this step of construction of the *decision tree* is simply a phase of evaluation of the respective merits of each technique and no decision is taken.

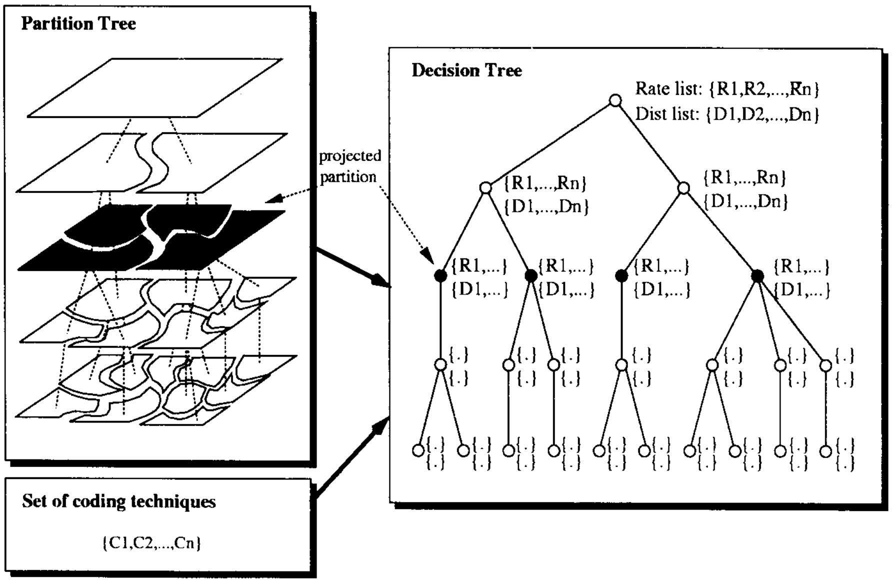

FIG. 3. — Construction of the decision tree.

Construction de l'arbre de décision.

The optimization relies on the technique discussed in [4, 7, 8]. It is a two step procedure :

- The first step is to make a local analysis and to compute, for each node, the Lagrangian for each coding technique. The technique giving the minimum Lagrangian is considered as the optimum one for this node and its Lagrangian is stored.
- The second step is to define the best partition. This can be done by a bottom-up analysis of the *decision tree*. The initialization is done by activating all the leaves of the tree. This means that we assume that the initial strategy is to code the frame with the finest partition of the *partition tree*. Then, starting from the lowest level, one checks if it is better to code the area represented by a set of children regions as a single region X or as a set of individual regions $\{x_i\}_i$ with $X = \cup x_i$. The selection of the best choice is done by comparing the Lagrangian of X with the sum of the Lagrangians of x_i. If the former is lower than the latter, the node corresponding to X is activated and the children nodes are desactivated. This procedure is iterated up to the root node. Note that, in order to use this approach the distortion should be additive over the regions. In our experiments, the squared error has been used, however, any additive measure can be used. Moreover, it should be noticed that the approach highly relies on the hierarchical structure of the regions in the *partition tree*.

At the end of the procedure, the best partition is defined by picking up all the regions corresponding to the activated nodes together with their corresponding best coding technique (defined during the first step of the algorithm).

II.2.2. Finding the best Lagrange parameter.

The definition of the optimum λ parameter can be done with a gradient search algorithm. The algorithm starts with one very high value λ_h (10^{20}) and one very low value $\lambda_l(0)$ of λ. For each value of λ, the optimization procedure described above is performed. This results in two coding strategies (partition and coding techniques) that should give rates R_h and R_l respectively below and above the budget. If none of these rates is close enough to the budget, a new Lagrange parameter is defined as $\lambda = (D_h - D_l)/(R_l - R_h)$. The procedure is iterated until the rate gets close enough to the budget [3].

II.2.3. Modification of the optimization strategy.

In this formulation of the optimization problem, the main parameter is the budget that is assumed to be given for each frame. Based on this budget, the algorithm finds the coding strategy that minimizes the distortion. In practice, this procedure creates a coded sequence with a variable quality. The same structure can be used to define a coding strategy leading to constant quality sequences. The only modification consists in defining a target distortion value for each frame and in inverting the role of D and R in the previous explanation. In this last case, the decision minimizes the coding cost to reach a given distortion.

However, in practice, an intermediate solution may be used. Indeed, working at a fixed cost per frame may produce some frames of very poor quality (scene changes, complex motion). Most of the time, it is more efficient to spend more bits for these frames so that the quality is not too low and that it may be possible to use these frames for the compensation of future frames. Therefore, the optimization can work basically on a fixed nominal budget, but a minimum signal to noise ratio for each frame can be defined. If this minimum signal to noise ratio is not reached with the nominal budget, the budget is progressively increased. The procedure is stopped when the decision has found the optimum strategy : 1) the distortion is minimal, 2) the budget is at least equal to the nominal budget and 3) the signal to noise ratio is above a given threshold.

II.3. Examples.

An example of *partition tree* creation and *decision* process is given in Figure 4. The image on the first row corresponds to the partition of the previous frame. This partition is projected and the projected partition can be seen in the center of the second row. This step defines the time evolution of the previously transmitted regions. Based on the projected partition, the *partition tree* is created : in the example of Figure 4, levels 1 and 2 are obtained by hierarchical segmentation following a spatial homogeneity criterion. Note in particular, how regions representing details of the face or of the background are introduced in the universe of regions. Levels 4 and 5 are created by merging regions with similar motion. Note here how background regions are merged because of their homogeneity in motion. The final partition is shown in the center of the lower row. In this partition, some regions are homogeneous in terms of gray level (regions corresponding to homogeneous part of the building) and others are homogeneous in motion (region corresponding to the face). Finally, the original as well as the resulting coded frames are shown in the lower row. The coding has been achieved using the algorithm described in [9] where more details can be found about the motion estimation/compensation and the coding techniques actually used.

In order to demonstrate the capability of the algorithm to cope with various types of sequences and scenarios, two different examples are presented now :

- The example of Figure 5 shows the frames of the sequence *foreman* coded at 42 kbit/s. This sequence is in QCIF format (176 × 144 pixels) and has been coded at 5 frames/s. The first row corresponds to the original frames number 0, 115 and 235 and the second row to the coded frames.
- In turn, the example of Figure 6 shows the frames of the sequence *children* coded at 320 kbit/s. This sequence is in QCIF format (176 × 144 pixels) and has been coded at 15 frames/s. The first row corresponds to the original frames number 6,100 and 192 and the second row to the coded frames.

The results of the bit allocation among the different types of information for both examples are summarized in Table I. Note that the *decision* algorithm has selected a quite different strategy for the two bit rates : for low bit rates almost 20% of the bit stream is devoted to the partition information whereas less than 10% is used for this type of information for higher bit rates. These figures illustrate one of the claims of *second*

TABLE I. — Bit allocation for two examples.

Allocation de bits pour deux exemples.

sequence	kbit/s	decision	motion	partition	texture
foreman	42	1.8 %	4.2 %	19.0 %	75.0 %
children	320	1.0 %	2.1 %	8.9 %	88.0 %

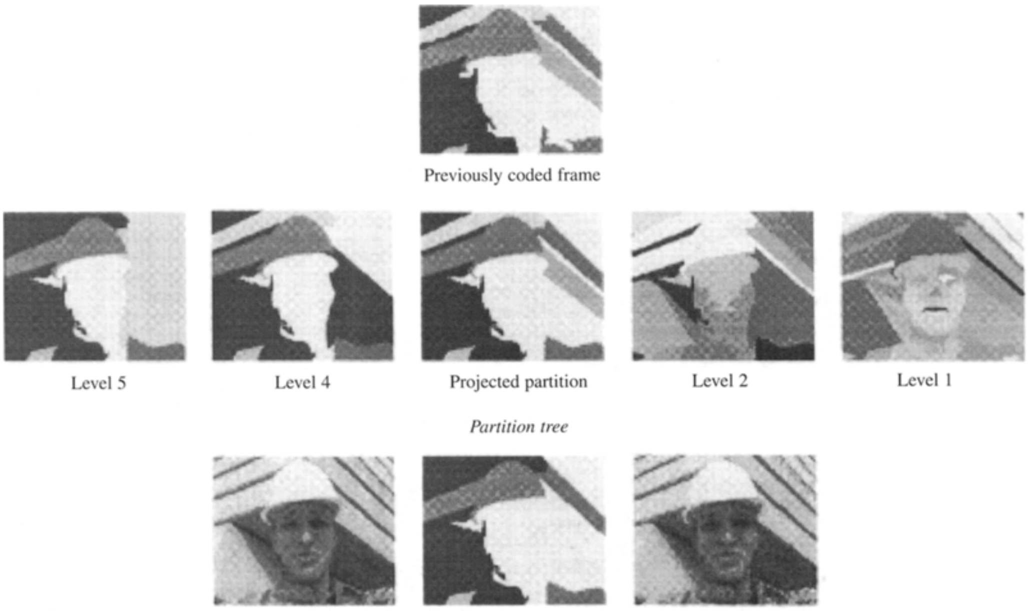

FIG. 4. — Example of inter-frame segmentation.

Exemple de segmentation inter-trame.

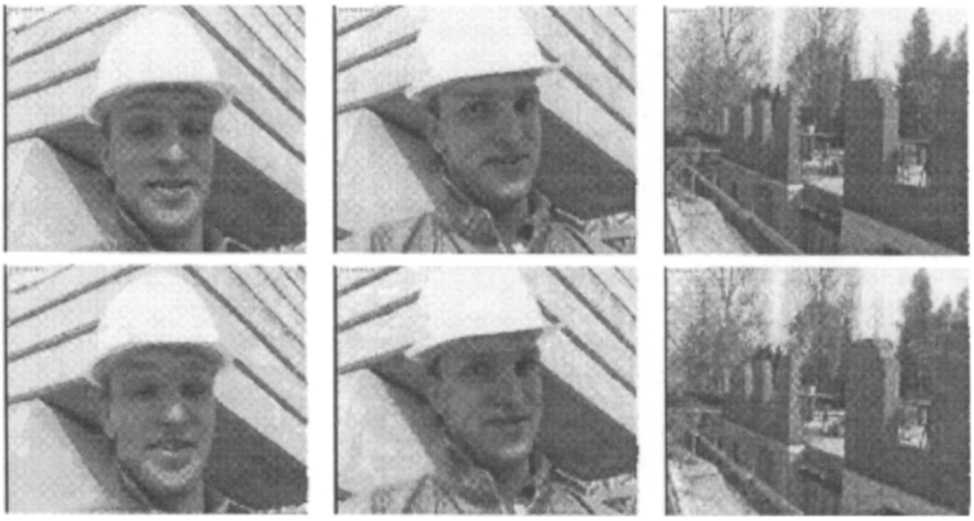

FIG. 5. — Coding results for the *foreman* sequence coded at 42 kbit/s and 5 frames/s ; first row : original frames 0, 115 and 235 ; second row : coded frames.

Résultats de codage pour la séquence foreman *codée à 42 kbit/s et 5 trames/s ; première ligne : trames originales 0, 115, 235 ; deuxième ligne : trames codées.*

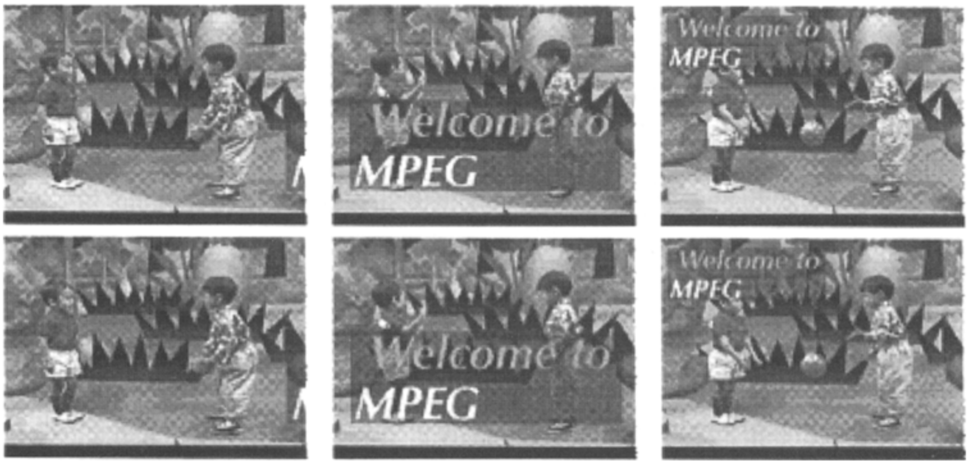

FIG. 6. — Coding results for the *children* sequence coded at 320 kbit/s and 15 frames/s ; first row : original frames 6, 100, 192 ; second row : coded frames.

Résultats de codage pour la séquence children *codée à 320 kbit/s et 15 trames/s ; première ligne : trames originales 6, 100, 192 ; deuxième ligne : trames codées.*

generation coding techniques stating that, at low bitrates, contour information becomes more important than at high bitrates.

III. SIMPLIFICATION OF THE OPTIMIZATION APPROACH

The approach described in Section II offers the advantage of optimally defining the segmentation and the coding strategy. However, for certain applications, it's computation cost may be prohibitive. Note that the complexity is not due to the optimization algorithm which is extremely fast but to the initial phase of computation of the rate and distortion for each region and all possible coding techniques. The purpose of this section is to describe several alternatives that can be used to simplify the optimization approach.

III.1. Reduction of the set of possible techniques.

As mentioned previously, the complexity of the algorithm is directly related to the size of the universe of regions and the number of coding techniques. The algorithm complexity can be severely decreased if the number of coding technique is limited.

It has to be noticed that, very often, while computing the rate and the distortion corresponding to a given coding technique on a region, one can *a priori* know that this coding technique is not going to be selected by the decision process because either it is too expansive in terms of bits or it results in a very high distortion. In fact, the optimization algorithm is going to look for techniques q_i that provide the optimum relation between the rate $R(q_i)$ and the distortion $D(q_i)$. This relation is given by the so-called Lagrange multiplier : $\lambda = D(q_i)/R(q_i)$ and is the same for all regions of the image. On an experimental basis, it can be observed that the optimum λ^* does not fluctuate very much from one frame to the next one.

Moreover, in practice, it is quite easy to order the set of coding techniques q_i in terms of their rate/distortion characteristics $(D(q_i)/R(q_i))$. As a result, by using the optimum λ^* parameter obtained for the previous frame, one can discard a very large number of coding techniques without having to actually compute their rate and distortion.

III.2. Optimization of the partition.

Following the simplification approach described in the previous section, one can define a decision process involving a single coding technique per region. In this case, the objective of the decision process is simply to select the best set of regions to form the partition out of the universe of regions.

A simple example in this line of simplification consists in assuming that all regions are coded by their mean value. In this case, the rate assigned to each region is equal to 8 bits and the distortion is the squared error with respect to the region mean. Note that in this case, the resulting partitions cannot be considered as being optimal for coding because the coding technique is too simple (the coding technique corresponds to very high values of $D(q_i)/R(q_i)$). However, the combination of the projection, the *partition tree* and the decision steps can be viewed as a sequence segmentation algorithm following a constant gray level homogeneity criterion.

A more realistic example for coding application would be to define one *typical* coding technique, that in average leads to the optimum relation $D(q_i)/R(q_i)$. Note that the definition of this *typical* coding technique depends heavily on the target bit rate.

Finally, let us mention that the concept of rate/distortion optimization can be used with other criteria. As seen previously, the distortion can be seen as a simple measure of the region gray level distribution such as the squared error with respect to the mean. The rate can be replaced by the number of regions or by the length of the contours of the region. These modifications lead to various segmentation approaches with several characteristics.

III.3. Simplification of the optimization problem.

Finally, let us mention a simplification of the optimization problem that has been used in [10, 1]. The general approach described in Section II consists in selecting the best regions in the universe of regions to construct the partition. A simpler approach involves the selection of one of the levels of the *partition tree* as optimum partition. The selection can be done with one of the criteria previously discussed : number of regions, length of the contours, squared error (assuming that each region is represented by its mean).

In [10], this criterion was used with a *partition tree* with two levels (the projected partition plus a partition obtained by contrast-oriented segmentation). The definition of the contrast level was done on the basis of the length of the regions contours.

IV. CONCLUSIONS

This paper has focussed on the relation between segmentation for coding and rate control. The efficiency of a segmentation-based coding scheme heavily relies on this step that defines how many and which regions have to be transmitted. This problem can be formulated as a rate/distortion problem leading to a *decision* process that not only controls the segmentation, but also defines the coding strategy to be used in each region. The *decision* can be viewed as a link between the traditional *analysis* and *coding* blocks of an encoder.

The decision relies on constrained optimization techniques involving dynamic programming algorithms. An important conclusion of the approach is that the objective of the *analysis* step is not to produce one single partition but a set of regions proposals (called universe of regions). In order to allow fast optimization, the set of proposals should be structured in a hierarchical way. The *partition tree* structure has been proposed for this purpose.

Finally, together with the general approach, several simplified versions of the segmentation control have been proposed and discussed.

Manuscrit reçu le 20 mai 1997.

REFERENCES

[1] Marcotegui (B.), Crespo (J.), Meyer (F.). Morphological segmentation using texture and coding cost. In I. Pitas, editor, *1995 IEEE Workshop on Nonlinear Signal and Image Processing*, Halkidiki, Greece (June 20-22, 1995), pp. 246-249.

[2] Marcotegui (B.), Meyer (F.). Bottom-up segmentation of image sequences for coding. *Ann. Télécommunic.* (1997), **52,** n° 7-8, pp. 397-407.

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

[4] ORTEGA (A.), RAMCHANDRAN (K.), VETTERLI (M.). Optimal buffer-constrained source quantization and fast approximations. *In Proc. IEEE Int. Symp. Circuits and Systems* (May 1992), **1**.

[5] PARDÀS (M.), SALEMBIER (P.). Segmentation of video sequences for partition tree generation. *Ann. Télécommunic.* (1997), **52,** n° 7-8, pp. 389-396.

[6] PARDÀS (M.), SALEMBIER (P.), MARQUÉS (F.), MORROS (R.). Partition tree for segmentation-based video coding. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'96*, Atlanta (GA), USA (May 1996), **IV,** pp. 1983-1986.

[7] RAMCHANDRAN (K.), VETTERLI (M.). Best wavelet packet bases in a rate-distorsion sense. *IEEE Trans. Image Processing* (Apr. 1993), **2,** n° 2, pp. 160-175.

[8] REUSENS (E.). Joint optimization of representation model and frame segmentation for generic video compression. *EURASIP Signal Processing* (Sep. 1995), **46,** n° 11, pp. 105-117.

[9] SALEMBIER (P.), MARQUÉS (F.), PARDÀS (M.), MORROS (R.), CORSET (I.), JEANNIN (S.), BOUCHARD (L.), MEYER (F.), MARCOTEGUI (B.). Segmentation-based video coding system allowing the manipulation of objects. *IEEE Trans. Circuits and Systems for Video Technology* (Feb. 1997), **7,** n° 1, pp. 60-74.

[10] SALEMBIER (P.), TORRES (L.), MEYER (F.), GU (C.). Region-based video coding using mathematical morphology. *Proc. IEEE (Invited paper)* (June 1995), **83,** n° 6, pp. 843-857.

[11] SHOHAM (Y.), GERSHO (A.). Efficient bit allocation for an arbitrary set of quantizers. *IEEE Trans. Acoustics, Speech and Signal Processing* (Sep. 1988), **36,** pp. 1445-1453.