in Proceedings International Conference on Digital Signal Processing, Cyprus, Jun. 1995.

Region Based Image and Video Compression

R. Leonardi *;
Signals & Communications Lab., Dept. of Electronics for Automation,
University of Brescia, Brescia, I-25123, Italy
Ph: +(39-30) 371-5434 Fax: +(39-30) 380-014
e-mail: leon@bsing.ing.unibs.it

Apr. 4, 1995

1. INTRODUCTION

It appears clear that images or image sequences cannot be simply described as sole realizations of stochastic processes. The structure or local organization of pixels does not obey only to statistical models. Images are formed from the projection of 3D world objects onto a plane, with portions of one object occluding portions of another one. The way these objects are arranged and the way they can be separated from one another has clearly to be an underlying characteristics of real images.

It is essential to keep the structure with which visual information is organized. To understand the 2-dimensional arrangement of objects, we introduce the *structuring element* keyword. What is a *structuring element*[1]? A structuring element is an organized set of pixels with certain local properties. For example, a contour segment is a structuring element: the local property being that it is defined by a series of pixels having maximum local gradient in one direction. An other example for a structuring element is a uniform textured region. A third example of a structuring element is an area of an image exhibiting some symmetry characteristics. More abstract models for structuring elements can be used by combining different less abstract structuring elements: In a video-telephony application, a human face can define such a structuring element. In this case a series of low-level abstraction structuring elements, i.e. contour segments, special regions (such as eyes, mouth), are combined with certain spatial constraints so as to form a human face.

The fundamental difference between structuring element based coding and waveform coding lies in 3 aspects:

1. A structuring element is identified by some relational properties between groups of pixels, which are basically spatial for image compression (spatio-temporal for video compression). These properties denote a local organization. A structural element property implies certain contraints on neighboring structuring elements (for example, there cannot be 2 neighboring contour lines, or a human face cannot contain 3 eyes).

2. Given a structuring element characteristics, specific coding methods can be developped, which must not change the structuring element characteristics (e.g., a contour segment cannot become a region), but which allow for a change in position, shape or appearance (luminance) of this element. In other words, structuring elements are invariant under compression. Moreover, the relationship that exists between two structuring elements is also invariant. By using these constraints, a priori information is made available, which automatically reduces the entropy of the source. In other words, we claim that coding visual information which constitutes a structuring element of an image implies some a priori knowledge on the distribution of neigboring structuring elements. It might be argued that this is also true given a certain waveform, but we make the conjecture here that this a priori knowledge is much richer because of the inherent characteristics of the structuring element has.

*This work was supported by the Italian Ministery for University and Scientific and Technological Research.

[1] We prefer to use here the terminology of *structuring element* rather than *object* to identify the structural property, rather than the 2D/3D characteristics of a physical object. As contour segments or group of feature points can define structuring elements, we prefer this terminology.

3. An image is represented as a combination of structuring elements of eventually different types, depending on how "easy" their identification is from the source signal (e.g., it may not be that simple to extract a face from an image sequence) and how "accurate" this representation is desired. These structuring elements are organized in a certain way one with respect to the other. When distortion is introduced, it is defined differently for each type of structuring element, according to the allowable changes in position, shape and appearance, so as to keep the spatial organization between structuring elements possible and the structuring element own characteristics unchanged. If parts of the source signal do not contain certain structuring elements (for example, non contour points in a contour based representation of the information), an algorithm can be developped to reconstruct with a certain fidelity the missing information.

In the next sections we first describe a possible classification of structural coding methods. We then focus on a simplified description of region based (or segmentation based) coding methods. The final section outlines the limitations of the region based coding techniques.

2. CLASSIFICATION OF STRUCTURAL CODING METHODS

Structural coding can be divided into 4 main categories:

1. Contour based representations (Synthetic highs based coding, Directional decomposition based coding, Sketch based representation)
2. Operator or Feature based representations (Anisotropic nonstationary predictive coding, Texture based coding, Fractal based coding, Symmetry based coding)
3. Segmentation based representations (Region-based coding, quadtree/octree based coding, Binary space partitioning tree based coding)
4. Model-based representations (application specific, e.g. head and shoulder model for video-telephony).

Recently, some of the more traditional techniques (DCT based coding, Vector quantization, wavelet or subband based coding, ...) have tried to use the non-stationary nature of visual information by adaptively quantizing each block according to its local properties. This kind of classification resembles to the idea of segmenting the visual information. An evident example of the success of such approaches can be seen in the definition of the *slice layer* and *M-quant* parameters used in the coding of moving pictures and associated audio proposed by the MPEG2 standard (see [1]).

Most structural coding methods require to perform an analysis of the local image structure (edge detection, uniform texture, symmetry detection, face extraction, ...) It therefore relates to the complex world of pattern recognition. At present, this has been the limiting factor for its success, given the ill-conditioned problem of pattern identification for arbitrarily complex natural scenes.

Table 1 summarizes the various techniques used in structural coding identifying for each one of them the type of structuring element used, the range of low bit-rates for which the particular method gives visually acceptable results, the type of artefacts created by the method as bit-rate is reduced, its use for video compression by extending the approach to the temporal domain.

Among all the *structural coding* methods, the most probably appealing is the segmentation based coding one as it relates the most to the physical nature of objects in a given scene. In this case, the structuring element is a region with the same attribute: uniform texture, uniform motion, uniform chrominance characteristics, ... The problem is to identify these regions with a certain accuracy in the original image, keeping in mind however that the objective function is the lowest possible rate for a given reconstruction quality. Coding will involve the description of the region shape and location on one hand, its appearance on the other hand (i.e. the luminance/chrominance information)[2].

[2]It is important to note that both shape, location and appearnce have strong ties between neighboring regions. In effect, once a region position and shape are known, part of the position and shape of its neighbors is known as well. In the case of appearance, the problem may seem less obvious, but given the segmentation criterion used, knowing one region appearance implies some knowledge on the apperance of its neighbors.

Method	Contour Rep.			Feature Rep.				Segmentation Rep.			Model Rep.	
	SH	DD	SC	APC	TC	FC	SC	AS	2^nT	BSPT	H&S	W
Str. El.	E	Dir. E	E.	Dir. E.	T.P.	C.T.	S.A.	A.R.	S.R.	C.P.R.	H.S.	F.
bpp	.4-1	.15	.05-.15	.27-.5	.002	.5-1	.5-.8	.08-.25	.2-.4	.1-.2	.05	.001
Vid. Cod.	N	N	Y	N	N	N	N	Y	Y	N	Y	Y

Table 1: Classification of structural coding

SH : Synthetic Highs
DD : Directional Decomposition
S : Sketch based representation
APC : Anisotropic Predictive Coding
TC : Texture Coding
FC : Fractal Coding
SC : Symmetry based Coding
AS : Arbitrary Segmentation
2^nT : 2^n Tree representation
BSPT : Binary Space Partitioning Tree representation
HS : Head-and-Shoulder model based coding
W : Wireframe model
Y : Yes

Str. El. : Structuring Element
E : Edge
Dir. E : Directional Edge

T.P. : Texture Primitive
C.T. : Contractive Transformation
S.A. : Symmetry Axis
A.R. : Arbitrary Region
S.R. : Square Region
C.P.R. : Convex Polygonal Region
H.S. : Head-and-Shoulder primitives
F. : Face primitives
N : No

3. THE SEGMENTATION PROBLEM

Up to now, only segmentations that define an exact partition of the image have been suggested[3]. In other words, if S is the image domain which is partitioned into N regions R_i, we have

$$\bigcup_{i=1}^{N} R_i = S \text{ and } \bigcap_{i=1}^{N} R_i = \emptyset \tag{1}$$

In order to achieve the best rate distortion objective, one needs to segment the image so as to keep the smallest possible rate for a given reconstruction quality of the original image. In other words, the goal is to obtain the $R(D)$ bound among the set of possible segmentations.

Given a certain segmentation which defines a partition $P_j = \{(R_1^{(j)}, ..., R_{N_j}^{(j)}) | \bigcup_{i=1}^{N_j} R_i^{(j)} = S \text{ and } \bigcap_{i=1}^{N_j} R_i^{(j)} = \emptyset\}$ of the image, the distortion associated with that partition $D(P_j)$ can be estimated by

$$D(P_j) = \sum_{i=1}^{N_j} \sum_{(x,y)\in R_i^{(j)}} d(I(x,y), I_i^{(j)}(x,y)) \tag{2}$$

where $d(a,b)$ is the distortion measure used, and $I_i^{(j)}(x,y)$ is the reconstructed signal in the i-th region $R_i^{(j)}$. Ideally one should choose a distortion measure which is a function of the partition itself ($d(a,b) \hat{=} d^{(j)}(a,b)$) as the perceived quality of the reconstruction may depend for a same pixel location on the segmentation itself.

On the other hand, the rate associated with P_j incorporates both the cost of encoding P_j, i.e. the partition itself, and the cost of obtaining the approximation $I_i^{(j)}(x,y)$ of the luminance for each region (or other attribute

[3] Segmenting an image into regions that may overlap each other may be attractive, but to the best of our knowledge this approach has never been used, probably as it is in contradiction with the objective of creating the segmentation by identifying areas of the image having a uniform attribute.

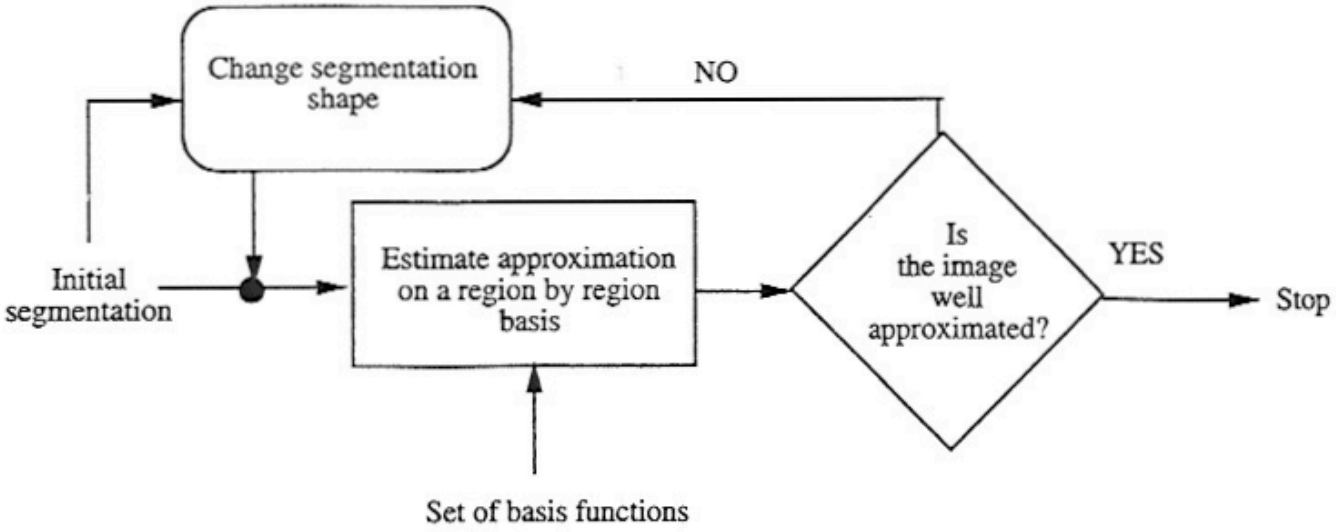

Figure 1: Adaptive segmentation principle

of interest, i.e. the appearance). With segmentation based models, it is clear that the underlying model is a non-continuous one. This is in accordance with the desire of separating non uniform areas of the image. Therefore, discontinuity may occur at region boundaries.

In a lossy coding scheme, quantization can lead to a simplified version of the of the partition P_j (i.e. another partition $\tilde{P}_j \hat{=} P_{j'}$). Quantization can also affect the approximation used in each region $R_i^{(j)}$.

In any region-based coding framework, the coding method for both the appearance and the segmentation shape description is highly dependent on the data structure used to obtain the segmentation. This makes it very difficult to study all possible segmentation/approximation pairs to seek the optimal code[4]. Given a certain segmentation structure (e.g. a binary tree), a coding method can be defined for describing both the partition and the approximation in a meaningful way.

In a sub-optimal approach, both the partition and the approximation shall be estimated jointly so as to minimize at each step the distortion measure D given in (2) (rather than finding the optimal rate-distortion bound). In practice, a change from an original segmentation will be performed so as to result in the slightest increase of the distortion measure, if the number of regions is decreased. Conversely, the new partition will be chosen so as to lead to a maximum decrease in distortion measure if the number of regions is increased. This process is called *adaptive segmentation* [2]. The diagram of Fig. 1 shows the Adaptive segmentation principle.

Given a certain partition of the original image domain R_0, and an initial parametric representation of each region $R_0^{(j)}$, how can the segmentation/approximation be changed so as to obtain a good rate-distortion? Typically a linear expansion of some appropriately chosen basis function $\{\psi_k(x,y); k = 1,..,r\}$ is used to obtain the approximate representation of each region. Given an "optimal" solution to the corresponding linear expansion for each region, the quality of the overall reconstructed image can be evaluated.

The optimality of this representation is often seeked in order to minimize some approximation criterion over the set of pixels in each region. The most often used criterion being the least square criterion. In this context an optimal solution can be found through a pseudo-inverse formulation. In other words, if there are $N_i^{(j)}$ pixels in a region $R_i^{(j)}$, and if the original *appearance* function is described by a matrix $I[m,n]$, its approximation $\tilde{I}_{ij}[m,n]$

[4]The coding strategy needs to be modified as the segmentation evolves, and it would be totally unrealistic to obtain an optimal operational rate-distortion bound.

is described for each pixel belonging to region $R_i^{(j)}$ by

$$\hat{I}_{ij}[m,n] = \sum_{k=1}^{r} u_{ijk}\psi_k(m,n) \qquad \forall (m,n) \in R_i^{(j)} \tag{3}$$

which if written in matrix form becomes

$$\hat{I}_{ij}[m,n] = \Psi^T[m,n].\mathbf{u}_{ij} \tag{4}$$

If all values $I_{ij}[m,n]$ are stacked into a vector $\mathbf{I}_{ij}$, the set of all approximated values can be expressed by the equation

$$\hat{\mathbf{I}}_{ij} = \mathbf{Z}_{ij}.\mathbf{u}_{ij} \tag{5}$$

where $\hat{\mathbf{I}}_{ij}$ represents the vector formed by stacking all values $\hat{I}_{ij}[m,n]$ and $\mathbf{Z}_{ij}$ is obtained by stacking all the vectors $\Psi^T[m,n]$.

Minimizing a square distortion measure over the entire region $R_i^{(j)}$ is equivalent to

$$\min_{\mathbf{u}_{ij}} d(I, \hat{I}_{ij})$$

where $d(.,.)$ is a square distance measure measured over all points in $R_i(j)$. This optimization problem leads to the well known pseudo-inverse solution given by[5]:

$$\mathbf{u}_{ij} = (\mathbf{Z}_{ij}^T\mathbf{Z}_{ij})^{-1}\mathbf{Z}_{ij}^T\mathbf{I}_{ij} \tag{6}$$

where $\mathbf{I}_{ij}$ is the vector obtained by stacking the set of all pixel values in $R_i^{(j)}$. After expansion the $\mathbf{Z}_{ij}^T\mathbf{Z}_{ij}$ becomes the $r \times r$ matrix

$$\mathbf{H}_{ij} = \begin{bmatrix} \sum_{\mathbf{x}\in R_i^{(j)}} \psi_1^2(\mathbf{x}) & \sum_{\mathbf{x}\in R_i^{(j)}} \psi_1(\mathbf{x})\psi_2(\mathbf{x}) & \dots & \sum_{\mathbf{x}\in R_i^{(j)}} \psi_1(\mathbf{x})\psi_r(\mathbf{x}) \\ \sum_{\mathbf{x}\in R_i^{(j)}} \psi_1(\mathbf{x})\psi_2(\mathbf{x}) & \sum_{\mathbf{x}\in R_i^{(j)}} \psi_2^2(\mathbf{x}) & \dots & \sum_{\mathbf{x}\in R_i^{(j)}} \psi_2(\mathbf{x})\psi_r(\mathbf{x}) \\ \sum_{\mathbf{x}\in R_i^{(j)}} \psi_1(\mathbf{x})\psi_r(\mathbf{x}) & \sum_{\mathbf{x}\in R_i^{(j)}} \psi_2(\mathbf{x})\psi_r(\mathbf{x}) & \dots & \sum_{\mathbf{x}\in R_i^{(j)}} \psi_r^2(\mathbf{x}) \end{bmatrix} \tag{7}$$

Starting from an initial segmentation P_0 made of N_0 disjoint regions $R_i^{(0)}$, it is now necessary to modify it to reach a new segmentation P_1 so as to reach a good rate-distortion tradeoff. The original distortion D_0 is obtained by summing all minimal distances $d(I, \hat{I}_{i0})$ related to all regions $R_i^{(0)}$. The corresponding rate needs to be estimated by obtaining a coded representation of both P_0 and the set of all appearance in the $R_i^{(0)}$ regions.

A better rate-distortion point can be obtained in several ways:

1. changing the shape of at least 2 regions without changing the number of regions forming the original partition, a *deformation*;
2. reducing or increasing the number of regions of the original partition, which not only creates or eliminates certain regions, but naturally involves a change in shape, and possibly connectivity.

We will not discuss the deformation approach as it falls beyond the scope of the methods presented hereafter, even though it may represent an attractive solution, for which interesting ad hoc solutions may be developed. Besides, a deformation can be simply considered as a limit case of the second approach as a deformation can be reached by adding then eliminating the same number of regions to specific parts of the original segmentation.

To add or remove regions on the other hand, we have to make the following considerations. The "ideal" partition should contain a number of regions directly related to the number of objects present in the original 3-D scene. These regions should match the apparent projections of the 3-D objects on the image plane. The degree of accuracy for describing the region appearance depends on the faithfulness of the representation. The larger the distortion the less adequate the correspondence between the approximated appearance and the original one.

[5]The solution is simply obtained by setting the derivative $\partial d/\partial \mathbf{u}$ to 0, and assumes that the matrix $\mathbf{Z}_{ij}^T\mathbf{Z}_{ij}$ is non singular

If P_0 is made of a very large number of regions, the "ideal" partition can be expected by successively merging different small regions into larger ones that will end-up matching the projections of the objects on the image plane. A good strategy to obtain a good segmentation is to merge at each step the two neighboring regions which are the most similar. A good degree of similarity between two regions is obtained by estimating the error resulting from the least square fit of the functional approximation over the merged configuration. Such error is clearly represented by[6] $d(I, \tilde{I}'_{kl,j+1})$ where

$$R_{kl}^{(j+1)} = R_k^{(j)} \cup R_l^{(j)} \tag{8}$$

Let us redefine such dissimilarity measure as $E_{k,l}$. It is possible this way to construct from partition P_j a Region Adjacency Graph (RAG), where each node corresponds to a region $R_i^{(j)}$ and links between nodes correspond to contiguous regions.

Each node in the RAG is assigned a model represented by a quantized version of the $\mathbf{u}_{ij}$ parameter vector, which defines the "optimal" least square approximation using the basis functions $\psi_k(x, y)$ for region $R_i^{(j)}$. Each branch linking regions R_k and R_l of the RAG is assigned a dissimilarity measure defined by $E_{k,l}$. All such dissimilarity measures are ranked in increasing order. If $E_{k*,l*}$ is the slightest dissimilarity measure, we set

$$R_{k*l*}^{(j+1)} = R_{k*}^{(j)} \cup R_{l*}^{(j)} \tag{9}$$

The algorithm can then be iterated by renaming $R_{k*l*}^{(j+1)}$ as $R_{\min(k*,l*)}^{(j+1)}$ and changing the links of the previous RAG, by linking the newly formed regions to the union of all neighbors of regions $R_k^{(j)}$ and $R_l^{(j)}$.

The algorithm stops so as not to exceed a target distortion. If a lossless coding scheme is used to encode the partition, the assigned rate will depend only on a quantization of the $\mathbf{u}_{ij}$ parameters. Compression can therefore be substantial if one can tolerate for the disappearance of small size objects.

4. CONCLUSION

In this paper, we quickly introduce the concept of strutural coding methods. We then described a possible classification of these techniques. The presentation was then centered on a simplified dscription of region based coding methods so as to obtain arbitrarily shaped regions.

The proposed strategy can easily be extended to video compression methods by either considering a spatio-temporal partition of the video sequence or by using the RAG formalism to obtain an accurate segmentation of the motion field.

5. REFERENCES

[1] *Coded Representation of Picture and Audio Information*, ISO-IEC/JTC1/SC29/WG11 MPEG technical document, 1993.

[2] R. Leonardi, *Segmentation Adaptative pour le Codage d'Images*, Ph.D. thesis 694, Dept. of Electrical Engineering, Swiss Federal Institute of Technology, Lausanne, Jun. 1987.

[6] We use here $\tilde{I}'_{kl,j+1}$ rather than $\tilde{I}_{kl,j+1}$ as it is a quantized version of the last one, obtained by quantizing the optimal parameter vector $\mathbf{u}_{kl,j+1}$.