SELECTIVE CODING BY FOCUS OF ATTENTION: A NEW TOOL TO ACHIEVE VLBR VIDEO CODING

invited paper

Eric Nguyen, Claude Labit
IRISA, Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
Tel: +33 99 84 72 60; fax: +33 99 84 71 71
e-mail: {nguyen,labit}@irisa.fr

ABSTRACT

Selective source coding is an essential part of very low bit rate (VLBR) image/video compression where a significant irrelevancy reduction has to be performed. In this paper, this reduction is described in the context of visual attention: the selection of relevant spatial information at the expense of other (non-relevant) information in order to maximize the efficiency of a particular visual communication task. We first give a general framework of selective coding. We then illustrate it with some examples of implementation using the generic wavelet representation as a stand-alone technique or for spatial encoding of the MC residuals in a MC-DPCM hybrid video coding scheme.

1 SELECTIVE CODING BY FOCUS OF ATTENTION

Visual attention is the ability to select elements of the visual field. Focalization, *i.e.* the concentration of perceptual resources on the selected elements of the scene, is its manifestation. The issue of selective resource allocation is the primary goal of VLBR image and video communications for which a significant irrelevancy reduction is necessary. It can be shown however that criterions for predicting the focus of attention of any observer looking at any scene are not well defined [5]. Consequently information regarding localization and saliency of the interesting parts of the scene to be coded should be given *a priori*. In the main applications of concern, general attributes of importance are actually assumed to be known: faces for videophone, motion-related events for video surveillance or pathology for telemedicine. This assumption refers to the task-driven approach of selective coding. The main steps in designing a selective compression scheme are illustrated hierarchically in Fig.1. This is the spectrum of compression mechanisms as inspired by the spectrum of attention mechanisms proposed in [9]. The bottom of the hierarchy (the selection of the task, application, and constraints) settles the specification of the coding problem. The source model (the image/video representation) is then selected to fit the required constraints. Finally, the R-D allocation

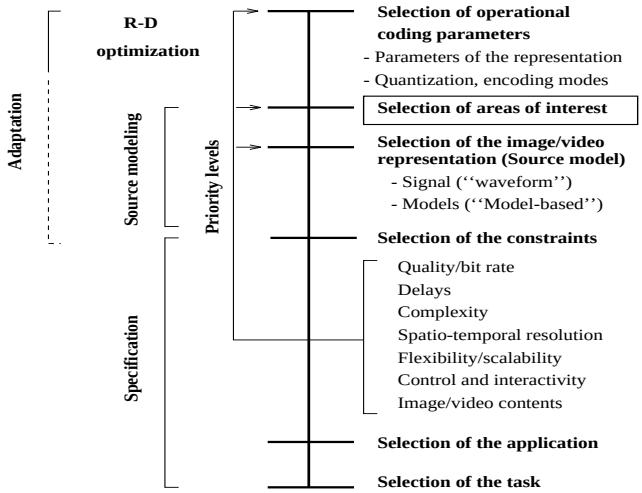

Figure 1: The spectrum of compression mechanisms.

performs the adaptation of the coding scheme by distributing the coding resources (rate R versus distortion D) to the different parts of the representation. This is the common framework of lossy compression shared by all image and video coding schemes. Selective coding by focus of attention refers to the inclusion of a spatial selection stage in the hierarchy (note that the selection of decoded quality of objects in the scene is also one of the functionalities suggested in the MPEG4 requirements [4]). This could be performed in several ways, either implicitly by a content-driven source modeling of specific scenes (such as in model-based coding for videophone applications), or explicitly by non-uniformly allocating the available R-D resources on a given image/video representation. In this paper we address the second way which involves the adaptation of state of the art coding techniques based on generic representations.

2 R-D framework

Fig.2 gives the general structure of a selective coding scheme. Selective coding is essentially based on two *a priori*:

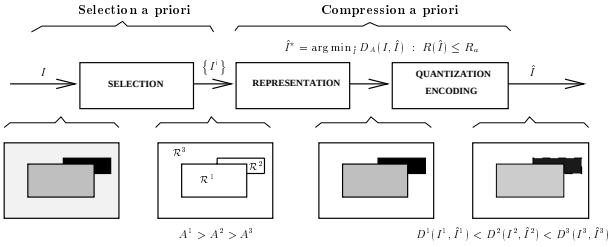

Figure 2: Selective compression.

1. The selection *a priori* which gives the location and importance of the different parts of the scene. It could be either a single location in what we call pin-point focalization, or a partition $\{\mathcal{R}^i\}$ linked with a hierarchy of interest $\{\mathcal{A}^i\}$ in what is usually called a region of interest (ROI) approach.

2. The compression *a priori* which gives the selected representation and the coding modes (selected parameters of the representation, quantization and binary encoding conditions) used to actually compress the visual information.

Selective coding is achieved by the choice of a suitable R-D objective function which should be minimized according to the usual R-D trade-off. Using a fixed rate framework, the aim is to non-uniformly distribute the quality of the spatial reconstruction under the constraint of a fixed global transmission rate R_a. The distortion measure $D_A(I,\hat{I})$ should both be global (for a global evaluation of the reconstructed frames $\hat{I}$) and include some local measures D^i to enable relative distortion allocation on the different parts of the scene) such that, at R-D optimality, the coded representation respect the given hierarchy of interest. Taking into account that visual communications concern essentially human observers, the distortion measure should further include some properties of the human visual system (HVS). Two approaches of selective compression can then be distinguished according to the two different ways of introducing HVS properties in coding systems: the psycho-visual way related to the pin-point focalization, and the conventional way related to the adaptation of state of the art compression techniques such as in ROI schemes. In the following, we briefly address some implementations of these two different approaches.

3 IMPLEMENTATIONS

Many implementations of selective compression are possible depending on the selection of the image/video representation. The necessary condition is that the representation should be localized in space. Furthermore, in order to translate the relative importance in terms of relative precision in the coding of the parameters of the representation, the representation of each part of the scene should be comparable; *i.e.* the same primitives should be used. In this study, we focus on generic representations and in particular on the widely used waveform basis functions primitives for spatial representation in transform, subband image and hybrid video coding. Wavelet/subband representation is used to illustrate both pin-point focalization and ROI video coding in a hybrid MC-DPCM scheme. These methods take advantage of the properties of space-frequency localization of the basis functions.

3.1 Pin-point focalization

In pin-point focalization, the selection *a priori* is given by a position of interest for each image in the sequence. Ideally, this position would be given by an eye-tracker recording the position of the eyes of the observer at the receiver. The aim of the compression is to simulate the peripheral contraction of spatial information performed by the retino-cortical sampling of the HVS. This could be done directly by re-sampling the digital images according to a logarithmic warping function of the retinal eccentricity r [1]. This function represents the integral of the deterministic $\frac{1}{r}$ approximation of the variation of the radial sampling density of the HVS. The direct approach introduces some local aliasing artifacts depending on the amplitude of the spatial compression. These artifacts are due to the fact that images are not generally locally band-limited according to the warping function used by the HVS. Consequently space-variant low-pass filtering should normally be applied before resampling. An alternative approach is to approximate the $\frac{1}{r}$ variation of the resolution by a piecewise constant function of resolution 2^{-j} where j is the resolution level of a discrete-time multiresolution representation of the signal (this was first proposed by Burt in a context of active vision [2] by using a Laplacian pyramid). This idea could apply for any subband representation and in particular a non-redundant multiresolution wavelet representation. In that case, it can be easily shown that the HVS-like representation of the signal could be obtained by discarding (thresholding) coefficients which are outside a "foveal" zone of a given radius for each subband of the pyramid. The scheme of the coding process is

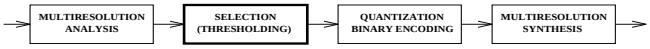

Figure 3: Coding process in pin-point focalization.

shown in Fig. 3. The compression mechanism can be divided into two steps. The wavelet representation of each frame is first thresholded according to the previous selection rule. This selection stage yields a spatial compression measured by the ratio g_s of the number of retained coefficients over the size of the image (which is the number of subband coefficients for a critically sampled subband representation). The remaining coefficients are then quantized using a classical subband quantization technique. In our implementation, intra-subband uni-

form threshold quantization and arithmetic memoryless encoding of quantization indexes are used. Fig. 4 shows

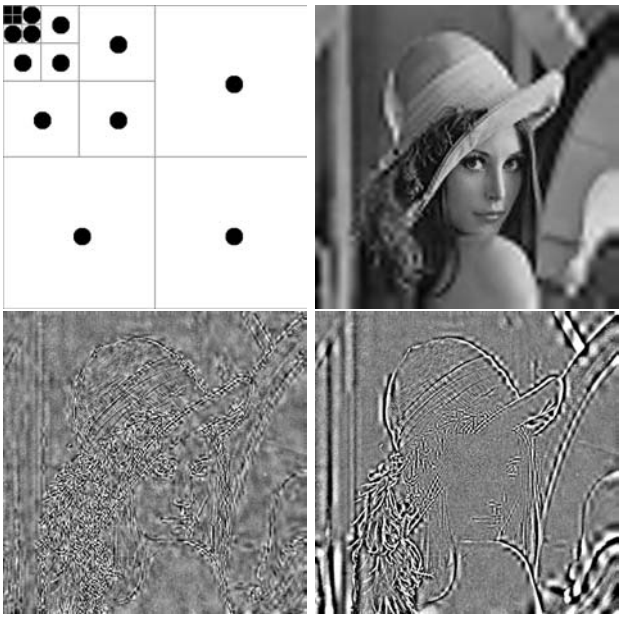

Figure 4: Up: Selected wavelet coefficients and reconstructed frame for a pin-point focalization centered in the Lena image. Bottom: errors images for the conventional and the selective coding approaches.

the selected coefficients, and the reconstructed frame for an overall compression ratio of 80 for the "Lena" image. Error images (enhanced) are also given in order to illustrate the difference between a selective coding approach (right) and a usual approach (left), *i.e.* w/o a selection stage, for the same overall compression ratio. The reconstruction quality obtained with the selective coding approach is judged to be globally enhanced thanks to the choice of a position of interest centered in the image. In this example, we chose a significant spatial compression ratio ($g_s = 28$) such that the actual bit allocation for the remaining coefficients is high (around $3bpp$). This causes the foveal zone to be quasi-perfectly reconstructed (for a PR filter bank) at the expense of the periphery.

Pin-point focalization is essentially based on a spatial compression matched to the one performed by the HVS. For graceful degradations, compression ratios are limited by the contraction performed by the HVS (typically in the region of 4 for a typical viewing distance). It's basically a single user compression scheme according to the fact that focus of attention strongly depends on the observer. The main limitation of a video coding scheme controlled by eyes movement is the delay introduced by the control as mentioned in [3]: the transmission time can delay the actual position fixed by the observer with the position taken into account in the coding process. As a conclusion, the applicability of the psycho-visual way for selective coding is limited to single user (assuming also that an eye-tracker is provided) and low delay transmission environments. In the following we address the "conventional" approach of selective coding where the selection stage is based on an *a priori* analysis.

3.2 ROI hybrid video coding

In standard hybrid video coding at very low bit rate (H.263), selective coding could be applied by local adaptation of the quantization of block DCT coefficients at the macro-bloc level depending on an analysis of the priority of each of these macro-blocks at the coder stage. The efficiency and potential functionalities of this scheme are however limited by the fixed block structure of the image partition. In a more flexible scheme, areas of interest should be defined on the basis of an arbitrary partition of each image for which to each region one could assign a measure of importance or a level of priority.

We have chosen an application where areas of interest are defined by a motion analysis [10]. We have proposed a method of hybrid video compression that takes into account a psycho-visual hypothesis: the region of interest is tracked by the eyes of the observer [7]. The

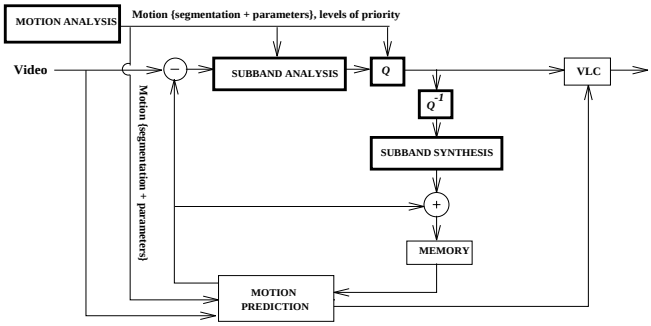

Figure 5: The ROI hybrid video coding scheme.

ROI hybrid video coding scheme is shown in fig. 5. The main differences with actual standards are the use of a general motion-based analysis (segmentation) [8] and the use of a region-based local adaptation of the quantization. Motion is indeed an essential primitive for the discrimination of areas of interest in applications such as surveillance. In our hybrid video coding scheme, motion information is used both for ROI selection and compression using motion compensation and suitable selective R-D allocation for the motion-compensated prediction errors (MCPE).

The motion parameters (affine model) are quantized according to the precision parameter (in residual displacement) of the motion analysis. The segmentation is lossless encoded in a contour basis using differential Freeman chaining and arithmetic encoding. Due to the simplicity of the motion-based segmentation obtained for a wide range of precision parameters and for usual VLBR video material, the bit rate of the motion analysis information {segmentation+parameters} remains acceptable (typically in the region of $0.01bpp$). In a fixed

rate environment, the remaining bits are allocated to each region by local adaptation of the MCPE quantization in the subband domain. Shape-adaptive region-based representation could be used. It can be shown however that boundary effects (such as those introduced by simple periodic extension at the regions boundaries) increase significantly the quantization noise. In our case we choose a hierarchical projection of the segmentation in the subband domain taking into account the spatial localization of the basis functions and the hierarchy of levels of interest. A weighted (in a region-subband sense) average l_2^2 distortion measure is used to specify the selective allocation. A psycho-visual *a priori* enables to weight region-subband contributions according to the frequency tuning of the HVS perception of moving objects. Assuming eyes movements tracking the ROI (the region of maximal priority) perform perfect motion-compensation, relative reconstruction quality is related to the degree of blurring effect introduced by the tuning of the visual velocity (that is the average velocity of regions relatively to the motion of the eyes) on the contrast sensitivity response of the HVS (see [7, 6] for details). Our approach has been validated for low bit

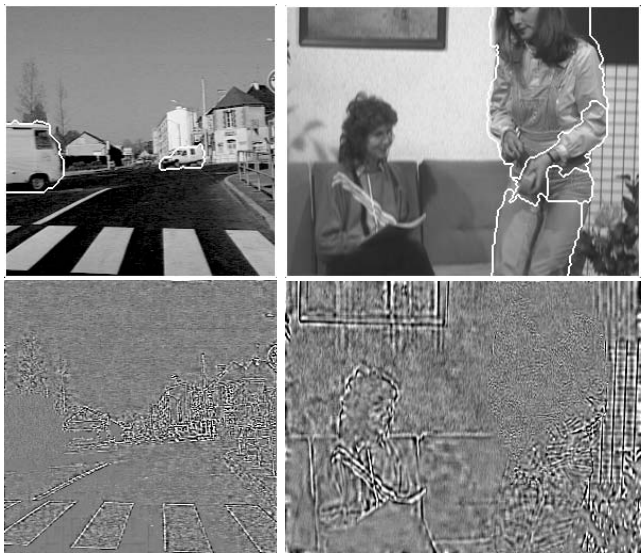

Figure 6: Examples of error distribution obtained in ROI hybrid video coding using motion-based analysis. Overall transmission rate is $0.1bpp$.

rates (around $0.1bpp$) on sequences of real images. Fig. 6 shows some examples of segmented frames and associated reconstructed errors (enhanced). Results have been subjectively compared with conventional approaches using global R-D allocation. In the ROI approach, the reconstruction quality of the ROI is enhanced in comparison with the quality of other regions which are blurred according to the motion-based focalization (this effect is clearly shown in the error images of Fig. 6). When the observer focuses on the ROI, distortions are less perceptible. The choice of a stationary subband representation along the time axis (insuring temporal consistency of the spatio-frequential representation) reduces the somewhat *flickering* effect obtained when using adaptive representations.

4 CONCLUSION

Selective compression enables to concentrate coding resources on the areas of interest at the expense of the other areas in the image, and thus obtain a significant gain in the issue of lossy compression at very low bit rate. Two methods have been proposed to illustrate such an approach using generic image/video representation and in particular the spatially localized subband (wavelet) representation using either the selection of the subband coefficients (pre-quantization) or the local adaptation of the quantization according to psycho-visual criterions. Extensions to other image/video representations and dynamic perceptual evaluations based on eye-tracker measurements are under investigations.

References

[1] Basu A. and Wiebe K.J. Videoconferencing using spatially varying sensing with multiple and moving foveae. In *Proc. IEEE Int. Conf. Pattern Recognition*, pages 30–34, 1994.

[2] Burt P.J. Attention Mechanisms for Vision in a Dynamic World. In *Proc. IEEE Int. Conf. Pattern Recognition*, pages 977–987, 1988.

[3] Girod B. Eyes Movements and Coding of Video Sequences. In *Proc. SPIE Conf. Visual Commun. Image Processing*, volume 1001, pages 398–405, 1988.

[4] MPEG4. Proposal Package Description, Revision 2. Technical Report ISO/IEC JTC1/SC29/WG11 N0937, 1995.

[5] Nguyen E. Compression sélective et focalisation visuelle: application au codage hybride de séquences d'images. PhD thesis, Université de Rennes I, December 1995.

[6] Nguyen E., Labit C. Selective coding using the AOI concept: application to VLBR video coding. *In preparation*, 1996.

[7] Nguyen E., Labit C. and Odobez J.M. A ROI approach for hybrid image sequence coding. In *Proc. IEEE Int. Conf. Image Processing*, pages 245–249, 1994.

[8] Odobez J.M. and Bouthemy P. MRF-Based motion segmentation exploiting a 2D motion model robust estimation. In *Proc. IEEE Int. Conf. Image Processing*, pages 628–632, 1995.

[9] Tsotsos J.K. On the Relative Complexity of Active vs. Passive Visual Search. *Int. J. Computer Vision*, 7(2):127–141, 1992.

[10] Tziritas G. and Labit C. Motion analysis for image sequence coding. Advances in Image Comunications, Vol. 4. Elsevier Science Publishers, July 1994.