A *ROI* APPROACH FOR HYBRID IMAGE SEQUENCE CODING

E. Nguyen, C. Labit and J-M. Odobez

IRISA/INRIA, Campus Universitaire de Beaulieu,
35042 Rennes Cedex, France
e-mail: nguyen@irisa.fr, labit@irisa.fr, odobez@irisa.fr

ABSTRACT

In this paper we present an approach of selective compression of an image sequence based on *a priori* selection of Region(s) Of Interest (ROI). This method relies on a given motion-based segmentation analysis. The problem of selective compression based on the concept of inhomogeneous spatial reconstruction quality is considered in the context of hybrid DPCM subband coding. Both spatial and frequency localization of the subband representation are explicitly used. Hierarchical compression is applied through adaptive quantization in the spatio-frequency domain. A weighted distortion metric is used to introduce both *a priori* and velocity-based visual masking.

1. INTRODUCTION

In the field of video sequence coding at low bit rates, transform or subband coding have been extensively studied simultaneously with Motion-Compensation (MC) prediction loops in inter-frame predictive coding schemes to obtain the best redundancy removal and the associated bit reduction, building so-called hybrid MC-DPCM coders [1]. Contrary to the usual situation of broadcast imagery transmission where it has to be considered that all parts of the processed pictures are of equal importance, it might be very useful to define some kind of "Region(s) Of Interest" for other communication services. Potential applications are very low bit rates encoding schemes and several image transmission applications where a spatially-constant reconstruction quality is not necessary ("Head and shoulders" scenes, remote surveillance video systems). Allowing coarser compression for non-relevant components of the signal can save bit rate. Alternately this can enhance local reconstruction quality for relevant components for a given bit rate. The idea of variable spatial reconstruction quality is feasible if *a priori* levels for reconstruction quality can be guided by the contents of the scene. This "focusing" approach yields two distinct algorithmic modules. The first one is the analysis-segmentation stage which provides useful information for the ROI selection based on *a priori* criteria (in our context, these are motion-based criteria). This first module is application-dependent. The second one is the lossy adaptive compression stage where the ROI-based approach generates suitable inhomogeneous spatial reconstruction quality for an overall bit rate or distortion allocation. In this study, motion-based analysis is performed prior to the compression stage. Motion information is then used both for ROI selection, compensation and selective compression in the hybrid DPCM coding scheme. Variable spatial reconstruction quality is obtained versus global optimization using a weighted Rate-Distortion ($R-D$) objective function in the subband domain.

2. MOTION-BASED ROI HYBRID DPCM CODER

The general scheme of the coder is shown in Fig. 1. In this scheme, the ROI selection is based on motion criterion and is either performed locally at the coder (active scheme) or controlled by a feedback loop according to the visual or side information sent to the receiver. At the coder, the motion analysis part for selecting the ROI could be done by the detection of independently moving objects [2] and the suitable allocation of different priority levels for these regions. However, we chose to perform a motion-based segmentation because on one hand it allows the use of more complex motion cues for the ROI selection [3], and on the other hand, the coding part can take advantage of this segmentation both for compensation tasks and psycho-visual analysis.

The goal of the motion-based analysis is to estimate the map of regions $\{R_k\}$ and their 2D motion descriptors $\left\{\Theta_k^{init}\right\}$ (typically affine motion models with respect to spatial coordinates) which best describes the motion activity between two images; this structure $\left\{R_k, \Theta_k^{init}\right\}$ gives a compact representation for coding facilities [4]. In our case, the motion models are estimated on each region of the projection of the previous segmentation map. In the next step, a statistical regularization (namely multiscale Markov Random Field (MRF)), which uses motion observations as well as their reliability, is performed to get the segmentation map. The third step consists in detecting the areas where the motion descriptors are not valid [2], and in creating new regions if necessary. For analysis purpose, the segmentation is computed with a pixel level precision, while for coding purpose, a coarser level in the multiscale MRF can be chosen according to the trade-off between the overhead for the segmentation information and the error prediction energy.

In the context of image sequence coding, some joint motion-based segmentation and compensation should be used in the DPCM loop. Moreover when dealing with hierarchical subband decomposition, the subband tree-structure could be used in coding the MC error frames as well as in estimating the motion in the image sequence [5]. However, since the previously encoded frames are corrupted by quantization errors, the motion-based segmentation can be biased and lead to bad boundaries. Thus, the segmentation $\{R_k\}$ obtained in the analysis module is kept unmodified in the coder. As shown in Fig. 1, the motion parameters $\left\{\Theta_k^{init}\right\}$ are refined (according to the previously encoded data) in the embedded prediction loop. Since typical iterative relaxation methods are used to compute the correction term $\{\Delta\Theta_k\}$, we expect the overall solutions to reach local minima close to the initial value of the unbiased open-loop motion estimation. In fact, may be due to the fact that we are dealing with region-based estimation, the experiments showed that $\{\Delta\Theta_k\}$ was nearly zero, resulting in no improvement of the MSE when using $\{\Theta_k\}$ rather than $\left\{\Theta_k^{init}\right\}$. Thus, taking open-loop motion descriptors $\left\{\Theta_k^{init}\right\}$ for the compensation lead to near optimal coding performances, and typical local MC artifacts due to quantization effects and uncorrelated with somewhat physical motion (which are known to be perceptually annoying) are then expected to be reduced.

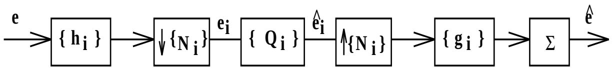
3. ROI APPROACH FOR MC-DPCM SUBBAND CODING

3.1. Problem statement : weighted l^2 distortion

Given the motion based information $\{\Theta_k,\ R_k\}$, the image sequence is compensated. Assuming that a hierarchy of *interest* has been derived from the ROI selection:

$$\{O_k\ :\ O_{ROI=0} > O_1 > \ldots > O_{R-1}\}\,, \tag{1}$$

subband coding of the prediction error (PE) signal e should generate inhomogeneous spatial reconstruction quality according to the previously mentioned hierarchy. As far as computational issue is concerned, transform, subband or wavelet coding is based on the classical analysis $\{h_i\}$ - synthesis $\{g_i\}$ FIR filter-banks multirate structure : where



$\{N_i\}$ denotes the subsampling/interpolation factors associated to each subband channel $i = 0, \ldots N-1$. Using discrete time formulation in the one-dimensional case, subband representation is given by the subband coefficients $\{e_i\}$ computed in the analysis part:

$$e_i(n) = \sum_j h_i(N_i n - j)e(j) =< e, \tilde{h}_i(j - N_i n) >_{l^2_{\mathbb{Z}}} \tag{2}$$

where $\tilde{h}_i$ denoting the time-reversed version of h_i is the underlying analysis subband basis function. The subband representation is localized both in space (on the spatial support $\tilde{L}_i(n)$ of $\tilde{h}_i(j-N_i n)$) and in frequency according to the frequency characteristics of each subband channels. In the synthesis part, the compressed signal is reconstructed using a quantized version $\{\hat{e}_i\}$ of the subband representation :

$$\hat{e}(n) = \sum_{i=0}^{N-1} \sum_j g_i(n - N_i j)\hat{e}_i(j) \tag{3}$$

Extension to the bi-dimensional case is performed using usual separable processing in both spatial directions. Classical l^2 distortion metric is used since decomposition and reconstruction are connected with unitary operators for paraunitary filter banks. Assuming independent encoding of each PE region signal $\{e_k\}$, we expect that dealing with region-based mean square errors will enable us to incorporate local rather than global metrics closer to human perception. Under these assumptions, overall distortion D and rate R are given by:

$$D = \sum_k \sum_{i_k} \eta_k D_{i_k}/N_{i_k} \text{ and } R = \sum_k \sum_{i_k} \eta_k R_{i_k}/N_{i_k} \tag{4}$$

where D_{i_k} and R_{i_k} are respectively distortion and rate of subband i of region k and η_k is the relative size of region k. Consider the fixed distortion allocation problem : $\min R\ :\ D \le D_0$. Under classical assumptions (asymptotic high resolution assumption leading to convex log variation of $R[D]$ functions, and same region-subband *pdf*'s) it can be shown that at $R-D$ optimality, relative contribution D_k^* of region k in the overall distortion is given by: $D_k^* = \eta_k D_0$. It means that optimal allocation will distribute the overall distortion according to the relative size of each region. Thus, in order to enhance relative reconstruction quality, *a priori* weighting factors $\{\Omega_k\}$ depending on relative relevance of each region should be introduced defining a region-based weighted l^2 metric :

$$D_w = \sum_k \sum_{i_k} \Omega_k \eta_k D_{i_k}/N_{i_k} \tag{5}$$

such that : $D_k^* = \eta_k D_0/\Omega_k$. The $\{\Omega_k\}$ can thus be considered as *quantitative* decimating factors in the relative distortion contributions of each region. At this point, only region-based spatial weighting are considered. The $\{\Omega_k\}$ have to be tuned according to quantitative *a priori* allocation in each region. In the following section, spectral weighting is introduced according to *psycho-visual* considerations.

3.2. Choice of the weighting factors

Since the resulting reconstructed frames are to be visually assessed, perceptual considerations about the Human Visual System (HVS) have to be taken into account. Taking advantage of the frequency nature of the subband representation, the frequency tuning of the HVS perception can be used. Weighting factors $\{\Omega_k\}$ are then considered as both spatial and frequency weighting factors $\{W_{k,i}\}$ and should reflect the relative importance of subband i of region k. Spectral weighting is essentially based on the use of experimental contrast sensitivity functions $W(f_x, f_y)$ which are first order approximations of the HVS response. When using motion-based analysis, the basic strategy is to adaptively process the region-based quantization depending on motion activity [6]. For spatial encoding of MC prediction errors, velocity-based spectral sensitivity response $W(f_x, f_y, v)$ are used [7] assuming foveal vision and isotropic response for either spatial frequency and velocity. The *visual* velocity v denotes the apparent velocity of the region relative to the focus of the eyes. It is well known that the perception of a moving object (and associated motion-compensated PE) heavily depends on whether or not the object is tracked by the eyes [8]. The difficult issue of locating where the viewer looks is here simply given by the ROI selection according to motion interpretation. We thus introduce a psycho-visual *a priori* which enables to weight region contributions according to the frequency tuning of the HVS perception of moving objects. Assuming eyes movements tracking the ROI perform perfect motion-compensation, object-based sensitivity response $W_k(f_x, f_y) = W(f_x, f_y, v_k - v_{ROI})$ can be defined. Average apparent speed of objects $\{v_k\}$ are computed thanks to the motion-based region descriptors. As the velocity of an object becomes greater, the peak frequency approaches zero frequency and the relative sensitivity decreases. Variable reconstruction quality will thus be considered as a degree of blurring effect introduced by the velocity-based frequency tuning in the visual masking phenomenon. Assuming additive white quantization noise in the subband, the l^2 weighting factors $\{W_{k,i}\}$ can be defined as "noise spectrum shaping" coefficients [9] depending on the frequency response of the synthesis filters g_{i_k} and on the velocity-based sensitivity function for each moving objects :

$$W_{i_k} \ \alpha \int_{\mathcal{B}_{ix}} \int_{\mathcal{B}_{iy}} W_k(f_x, f_y) \ |G_{i_k}|^2 \, df_x df_y \qquad (6)$$

The use of spectral weighting factors enables to exploit the subjective redundancy of the HVS and experimentally gives rise to a slight smoothing effect thus reducing usual artifacts in subband coding at low rates. Spatial contribution $\{S_k\}$ can also be introduced. In particular, luminance masking effect can be taken into account and quantitative *a priori* factors can be used to enhance potentially relevant non-tracked objects. Normalization factors can also be considered to equalize relative contribution of PE signal variance and size for each region in order to deal with homogeneous relative normalized distortion. Finally, using $\Omega_k = S_k W_{i_k}$ we define the spatio-frequential weighted l^2 metric :

$$D_w = \sum_k S_k \eta_k \sum_{i_k} W_{i_k} D_{i_k} / N_{i_k} \qquad (7)$$

It stands for a *given* choice of the subband representation for each region through the use of the weighting factors W_{i_k}.

3.3. R-D optimization

Solving the $R - D$ constrained optimization problem relies on the choice of effective subband representation for each region $\{e_i\}_k$. Adaptive space-frequency tiling has been proposed [10] in order to take into account non-stationarity of the input signal to be coded. Adaptation is expected to improve performances in the case of predictive coding since any motion-estimation scheme generates highly non-stationary unpredictable MC errors. In the particular case of the l^2 metric and for orthonormal wavelet packets (WP) subband representation, the constructive WP tree structure leads to a fast pruning algorithm which enables to find the best basis in the $R - D$ sense [11]. However, the weighted l^2 metric (7) depends on the particular choice of a subband topology and prevents any direct comparison between different WP basis (i.e. the weighted norm l_w^2 is no longer conserved in the recursive construction of the WP tree). Eventually, the subband structure should be first adapted using only spatially weighted distortion ($\Omega_k = S_k$) and then optimal $R-D$ allocation should be performed on the resulting region representations. In order to reduce the computational complexity we adopted a fixed subband representation (multiresolution structure). The use of a multiresolution structure (most of the time suboptimal in the $R - D$ sense with l^2 metric for PE signal) yields reasonable visual reconstruction quality and is generally assumed to give a better match to low level vision mechanisms. Furthermore, we choose this overall subband representation for the whole PE frame (and thus for each region). The quantization is adapted locally thus exploiting explicitly the spatial localization in the overall subband representation. It leads to assume the following facts :

- the reconstructed frame being assessed globally, relative reconstruction quality with respect to the ROI should be assessed according to the precision of quantization for the same spatio-frequential entities, i.e. for the *same* spatio-frequential representation.
- thanks to the overlapping nature of the subband basis at region boundaries, an overall subband representation reduces the problem of arbitrary blocking artifacts which could appear for independent region processing at low rates (involving spatial partitioning by the use of appropriate extension or boundary filters [10]).

When using a global subband representation, the segmentation map $\{R_k\}$ should be appropriately scaled and projected in the spatio-frequency domain. Though reducing boundary effects, the overlapping nature of the representation makes the spatial region information spread among subband coefficients in the neighbourhood of region boundaries. In order to preserve the hierarchy of relevant spatial information in the decimation process, we choose a prioritized projection of the segmentation in the subband domain. More precisely, labeling of subband coefficients $e_i(n)$ is given according to the most relevant region contained in the support of the equivalent basis function in the pel domain, ie:

$$lab\,[e_i(n)] = Arg \max_{lab \in \tilde{L}_i(n)} O_{lab} \qquad (8)$$

Notice that attention should be paid on the choice of the filter banks for accurate space-frequency localization. In all cases, spatial localization is given by the spatial support of the underlying analysis subband basis functions $\tilde{L}_i(n)$. However exact localization is not required in the coding process since the above hierarchical labeling method insures a good reconstruction quality in the area including the ROI. Furthermore the segmentation information is transmitted as side information and can be used for analysis purpose.

Given a discrete choice for quantization parameterization in each subband-region $\{Q\} = \{q_{i,k}\}$, optimal $R - D$ allocation can be derived numerically through classical unconstrained Lagrange multiplier formulation [11]. Denote by R_a the overall admissible rate of the transmission channel, R_o the overhead information including the coded segmentation map, motion and quantization parameters and

R_b the resulting budget rate $R_b = R_a - R_o$ for the PE signal. For fixed rate allocation problem, the optimal quantization $\left\{q^*_{i,k}\right\}$ associated with the optimal operating point of slope λ^* on the convex hull of the overall $R-D$ discrete functional is given by the maximization of the biased overall Lagrangian cost $F(\lambda)$:

$$\lambda^* = Arg \max_{\lambda \geq 0} F(\lambda) \qquad (9)$$

$$F(\lambda) = \sum_k \sum_i \eta_{i_k} / N_{i_k} \min_{q_{i_k}} \left[S_k W_{i_k} D_{i_k} + \lambda R_{i_k} \right] - \lambda R_b$$

Analytical or numerical approximations could be derived assuming either $R-D$ convexity assumption (high resolution hypothesis) and/or accurate subband pdf's modeling such as generalized gaussian distribution [12, 5].

4. EXPERIMENTAL RESULTS

The above described scheme has been applied for a typical traffic control scene (see Fig. 2). The van which undergoes translational motion from left to right is our ROI. Levels of interest are given relatively to it. Simple uniform threshold quantization is used on wavelet coefficients obtained using Daubechies filters of length 4 [13]. The wavelet tree has been limited to three levels depth. The first frame was intra-coded at $0.5bpp$. In this example, overhead information (segmentation map and region-based motion descriptors) is considered as negligible (less than $0.04bpp$). We show an example of an inter-coded reconstructed frame for a given budget rate $R_b = 0.1bpp$ (entropy rate) using the motion adaptive region-based weighted metric (only the velocity-based weighting factors are considered). Results are subjectively compared with non-weighted l^2 metric for the same rate using the adaptive Wavelet Packets algorithm [11] on the same PE signal. In the ROI approach, the quality of the ROI reconstruction is enhanced in comparison with the perception of non-relevant regions (background) which are blurred according to the velocity-based visual masking (this effect is clearly shown in the error frame of Fig. 2). This induces the observer to naturally focus on the ROI. In this case, background distortions are hardly perceptible. The choice of a fixed subband representation along the time axis (insuring temporal consistency of the spatio-frequential representation) reduces the somewhat *flickering* effect obtained when using the adaptive WP representation. The reconstructed sequence appears to be smoothed and is perceptually more pleasant even though the representation is suboptimal in the l^2 sense.

5. CONCLUSION AND FUTURE WORK

A ROI hybrid DPCM subband image sequence coding scheme has been presented. In this approach the motion-based analysis is made independent of the coding stage. A general motion-based segmentation is used to compensate the image sequence signal and to arbitrarily select the ROI for hierarchical compression in the coding process. Side information (motion-based descriptors and segmentation) and prediction error signal are thus made independent of each other. Further studies should include both stages in a global $R-D$ allocation problem. Motion-based analysis and interpretation should take into account region-based quantization if embedded in the DPCM loop. In the coding part, a fixed overall subband representation has been used taking into account both spatial and frequential localizations. Extensive experimental studies should compare adaptive subband techniques taking into account the increase in computational complexity.

6. REFERENCES

[1] GHARAVI H. – Subband coding of video signals. – In *Subband image coding* (Kluwer Academic Press), pp. 229–271, 1991.

[2] ODOBEZ J.M and BOUTHEMY P. – Detection of Multiple Objects using Multiscale MRF with Camera Motion Compensation. – In *Proc. of ICIP'94.*, November 1994.

[3] BOUTHEMY P. and FRANCOIS E. – Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence. – *Int. J. Comp. Vision*, Vol. 10, No 2: pp. 157–182, 1993.

[4] NICOLAS H. and LABIT C. – Region-based motion estimation using deterministic relaxation for image sequence coding. – In *Proc. of ICASSP'92.*, Vol. 5, March 1992.

[5] NAVEEN T. and WOODS J.W – Motion compensated multiresolution transmission of high definition video. – *IEEE Trans. Circ. Syst. for Video tech.*, Vol. 4, No 1: pp. 29–41, February 1994.

[6] LI N. and al. – Using subjective redundancy for DCT coding of moving images. – In *Proc. of SPIE-VCIP 93.*, Vol. 2094, pages 1571–1580, 1993.

[7] KELLY D.H. – Motion and vision II. Stabilized spatio-temporal threshold surface – *J. Opt. Soc. Am*, Vol. 69, No 2: pp. 1340–1349, 1979.

[8] GIROD B. – Eyes Movements and Coding of Video sequences. – In *Proc. of SPIE-VCIP 88.*, Vol. 1001, pp. 398–405, 1988.

[9] VANDENDORPE L. – Optimized quantization for image subband coding. – *Signal Process.: Image Comm.*, Vol. 4, No 1: pp. 65–79, 1991.

[10] HERLEY C and al. – Tiling of the time-frequency plane: construction of arbitrary orthogonal bases and fast tiling algorithms – *IEEE Trans. on Signal Process.*, Vol. 41, No 12: pp. 3341–3359, December 1993.

[11] RAMCHANDRAN K. and VETTERLI M. – Best Wavelet Packet Bases in a Rate-Distortion Sense. – *IEEE Trans. Image Process.*, Vol. 2, No 2: pp. 160–174, April 1993.

[12] NGUYEN E. and LABIT C. – Quantitative definition of psycho-visual weighting matrices for adaptive scalar quantization in subband image coding. – In *Proc. of IEEE-IMDSP Workshop.*, September 1993.

[13] DAUBECHIES I. – Orthonormal bases of compactly supported wavelets – *Comm. on Pure Appl. Math.*, Vol. XLI, pp. 909–996, 1988.