SELECTIVE H.264 VIDEO CODING BASED ON A SALIENCY MAP

O. Le Meur [a,b] P. Le Callet [b] D. Barba [b]

[a] *THOMSON R&D France, 1 Avenue Belle Fontaine, 35511 Cesson-Sevigne, France*

[b] *IRCCyN UMR n6597 CNRS Ecole Polytechnique de l'Universite de Nantes rue Christian Pauc, La Chantrerie, 44306 Nantes, France*

Abstract

The demand in modern multimedia data transmission is continually increasing. New compression standard, such as the recent H.264/MPEG-4 AVC video coding standard, drastically improves the compression ratio. This higher compression ratio is required because the amount of multimedia data to transmit increases and that the perceived quality expected by the end user is not lessened. A complementary solution to maintain or to improve the perceived video quality is to selectively compress the information: perceptually non-important areas are subjected to higher compression than more relevant parts of the data. Therefore, the perceived quality can be heighten without increasing the bit rate. In this paper, we propose a new two-pass based coding framework driven by a saliency map. A saliency map is a 2D topographic map encoding the capacity of each pixel to attract the visual attention. This map stems from a computational model simulating the human visual attention. A selective compression is achieved by an a-priori knowledge of the spatial locations of the visually important areas. First, the proposed algorithm determines a quantization step for each macroblock according to a given bit rate (inferior to the global transmission rate). The second pass allocates the remaining bit budget by using the saliency map. Experiments based on the H.264 coding scheme were conducted leading to an undeniably objective quality improvement of the most interesting areas.

Key words: Selective compression, H.264, visual attention, saliency map
PACS:

1 Introduction

Encoding a video sequence at low bit-rate with good quality remains a major challenge in video coding research. Classical encoding methods seeking to

uniformly distribute the coding resources make this problem even more difficult to solve. However, it is commonly accepted that each frame of a video sequence has only a few limited regions, which are perfectly examined by the gaze of observers. Therefore, it may not be necessary to encode the totality of a video frame with uniform quality. In others words, the rationale is that a high fidelity reconstruction is only desirable on areas having the capacity to attract the visual attention. This type of encoding scheme is called selective compression or compression with Region of Interest (RoI). An interesting study conducted by A. Bradley [1] has shown that the perceived quality could be significantly enhanced if the size of the RoI is small and if the target bit rate is low enough to produce visible compression artifacts on the RoI. Nevertheless, these non conventional compression methods are not widely used. The major problem concerns the detection of the spatial locations of the regions of interest.

Currently, most of the proposed selective compression methods make the assumption that an a-priori knowledge about the spatial locations of the RoI is available. For example, Leontaris et al. [10] proposed a solution to allocate more bit rate to the visually relevant areas. In the same spirit of the work by Yang [18], the bit budget allocated to the RoI is increased by relaxing the bit rate constraint in a rate-distortion optimization. The relaxation is generally obtained by including a relaxation factor in the rate-distortion function. This factor ranges from 0 to 1 (0 for macroblocks belonging to the RoI).

Others methods are dedicated to particular applications (medical imaging, video conferencing) in which the numerous constraints allow to quite easily detect the RoI. In the context of TV broadcasting, there is currently no usable selective compression scheme based on an automatically extraction method of the RoI, useful for the TV broadcasters.

Recently, key advances have been achieved in computer vision algorithm, and more especially in the visual attention modeling. Algorithmic methods have been proposed to automatically select regions that are of interest. Computational models simulating the properties of Human Visual System (HVS) provide the best results. They aim at building a saliency map that is a 2D topographic map encoding the capability of each pixel to attract the visual attention. The most famous model was developed by L. Itti [7]. This first model raised a lot of interest, leading to several studies [19,16,2].

This paper describes a selective H.264 compression scheme driven by a spatio-temporal saliency map. As the design of the model is not the main topic of this contribution, the key points of the computational model of the selective visual attention are only briefly described in the section II. Readers could find more details both on the design of the model and its performances in references [14,15]. The selective compression scheme is described in the section III. The aim is to heighten the quality over the areas that have the highest saliency values. Section IV presents and examines the performances of the selective coding compared to a conventional coding. Finally, some conclusions are drawn in section V.

2 Computational model of selective visual attention

The process allowing to build the saliency maps are detailed in the following sections. The major step are just reminded because several papers have already described the model used in this study [14,15]. Figure 1 gives the flow chart of the proposed model.

2.1 Spatial saliency map

According to an important psychovisual backing, the spatial model consists of three sequential steps: visibility, perception and finally perceptual grouping step.

The visibility step simulates the limited sensitivity of the HVS. Despite the seemingly complex mechanisms underlying the human vision, the visual system is not able to perceive all information present in the visual field with the same accuracy. To take into account these intrinsic limitations, the visibility step includes the following set of basic mechanisms entirely identified and validated from psychophysic experiments:

(1) a transformation of the RGB component into the Krauskopf color space composed of the cardinal direction A (achromatic), Cr_1 (red and green opponent component) and Cr_2 (the blue and yellow opponent component) is achieved.
(2) the early visual features extraction achieved by a perceptual channel decomposition consists in splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation. This decomposition is applied on each of the three perceptual components. From psychophysic experiments, A. Senane [17] shows that psychovisual spatial frequency partitioning for the achromatic component leads to 17 psychovisual channels in standard TV viewing conditions while only 5 channels are obtained for chromatic components. Each resulting subband or channel may be regarded as the neural image corresponding to a population of visual cells tuned to a range of spatial frequency and to a particular orientation.
(3) Contrast Sensitivity Functions (CSF) have been widely used to measure the visibility of natural images components. In fact, these components can be described by a set of sinusoidal Fourier and their amplitude. The visibility of a specific component can be assessed by applying a CSF in the frequency domain. When the amplitude of a frequency component is greater than a threshold, the frequency component is perceptible. This threshold is called the visibility threshold, and its inverse defines the value

of the CSF at this spatial frequency. In the approach presented here, CSFs are applied to each components (A,Cr_1,Cr_2). A 2D anisotropic CSF designed by S. Daly is applied on the achromatic component [3]. The CSFs of the two color visual components Cr_1 and Cr_2 are modeled using sinusoidal color gratings [9]. These two last CSFs show that the human eye is more sensitive to chromatic components with frequencies up to 4-5 cpd. Sensitivity rolls off at both higher and lower frequencies.

(4) masking effect refers to the modification of the Differential Visibility Threshold (DVT) CT_0 of a stimulus due to the influences of the context, called the masking signal [13]. The value CT_0 of the DVT is modified into a value called CT by the masking effect. This modification is simply given by the relation $CT = CT_0 \times T$. When $T \geq 1$, the threshold increases meaning that there is a masking effect. When $0 \leq T \leq 1$, the threshold decrease corresponding to a pedestal effect. The visibility of the stimulus is increased. Most of the time, psychophysics experiments based on the detection of simple signals (such as sinusoidal patterns) are used to determine an analytic expression for the visual masking. It is obvious that this is a strong simplification with regard to the intrinsic complexity of natural pictures. Nevertheless, numerous applications (watermarking, video quality assessment) are built around such principles with interesting results. A psychovisual subband decomposition performed on each component is mandatory in order to tackle the three types of visual masking: intra-channel masking, inter-channel masking and inter-component masking. These types of masking and the analytic model we use are fully defined in [12].

These first steps built a spatial psychovisual space in which all the spatial information of the image are coherently normalized regarding the DVT.

The second part of the spatial model deals with perception. It aims at determining from the psychovisual space a description useful to the viewers and not cluttered with irrelevant information. To deal with a large amount of visual information, the visual system has attentional mechanisms for selecting relevant areas and for reducing the redundancy of the incoming visual information. In order to form an economical representation of the visual world, the particular oriented center/surround organization of the cortical cells is really important. For instance, center/surround organizations imply that visual cells are insensitive to uniform illumination. The responses of such cells are efficiently simulated by a difference-of-Gaussian function. Two relevant structural descriptions (one for the achromatic component and one for the chromatic components) are then built.

Perceptual grouping refers to the human visual ability to group and bind visual

features to organize a meaningful higher-level structure. There are numerous mechanisms involved in the perceptual grouping. One of the most common is the facilitative interactions that has been reported in numerous studies. In most cases, these interactions appear outside the Classical Receptive Field (CRF) along the preferred orientation axis and are maximal when center and surround stimuli are iso-oriented and co-aligned [8] (due to the long-range horizontal connections). In other words, the activity of cells is enhanced when the stimuli within the CRF and a stimuli within the surrounding area are bound to form a contour. This facilitative interaction is usually termed contour enhancement or contour grouping. Most of the recent computational models are based on the Gestalt principles of colinearity and proximity [5].
A achromatic (respectively chromatic) saliency map is then obtained by linearly combining the achromatic (respectively chromatic) subbands. The combination of different saliency maps into an unique map is really difficult. When several visual dimensions are considered, such process is mandatory in order to compute a single measure of interest for each location. The major problem concerns the merger of features stemming from different visual dimensions and having different dynamic ranges. A basic fusion algorithm is currently used.

In [14], we have shown that the spatial model provides a linear correlation coefficient close to 0.72 between the output of the model and the one stemming from eye-tracking experiments. This model can predict locations that are fixated by human observers better than three others models (an uniform, a gaussian and the model of L. Itti) on still images.

2.2 Temporal saliency map

Figure 1 gives the synoptic of the proposed algorithm used to build the temporal saliency map. The temporal saliency map is based on the fact the visual attention is attracted by motion contrast. Such contrast is deduced from the local and the global motion.
Motion estimation plays a very important role in the construction of the saliency map. Apparent motion is first computed between two successive frames using a hierarchical block matching method. In general, the hierarchical decomposition is performed by a dyadic Gaussian pyramid: the input image is first filtered by a 2D separable filter and then subsampled (horizontally and vertically by a factor of two). This process is iteratively applied to build up each level of the Gaussian pyramid. Here, we take advantage of the perceptual channel decomposition performed during the first steps of the spatial saliency map creation.
Two neighboring pictures are used to form two pyramids. $\overrightarrow{V}_n^i$ denotes the motion field for the n^{th} frame, at the i^{th} level of the pyramid. At the lowest

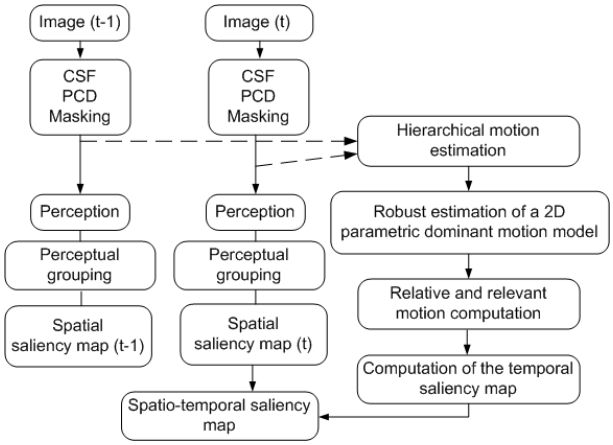

Fig. 1. Architecture of the proposed model: CSF (Contrast Sensitivity Function), PCD (Perceptual Channel Decomposition).

resolution, the motion vector providing the smallest sum of absolute difference is kept, up sampled and transmitted to the next higher resolution. Refinement algorithm and decision are used to form the final motion field $\overrightarrow{V}_{local}$.

In order to detect the temporal conspicuous areas of a video sequence, the motion inherent to the camera is first canceled. Assuming that the dominant motion is due to the camera, the global transformation between successive images based on a previous motion fields is estimated. The displacement $\overrightarrow{V}_{\Theta}(s)$, at a pixel site s related to a motion model parametrized by Θ is given by a 2D affine motion model. The six affine parameters are computed with a popular robust technique based on the M-estimators.
Finally, the relative motion is computed from the knowledge of the apparent dominant displacement $\overrightarrow{V}_{\Theta}$ and of the local displacement $\overrightarrow{V}_{local}$ for each pixel s. The relative motion $\overrightarrow{V}_{relative}$ is given by the relatio (1):

$$\overrightarrow{V}_{relative}(s) = \overrightarrow{V}_{\Theta}(s) - \overrightarrow{V}_{local}(s) \tag{1}$$

As the perception of a moving object heavily depends on whether or not the object is tracked by the eyes, the maximal pursuit displacement capability of the eyes ha sto considered. The relative displacement greater than the maximal pursuit displacement is discarded. For video, this value belongs to the range 8 to 10 deg/s.
The relevance degree of a relative motion also depends on the average amount of the relative displacement computed across the picture. For example, a high relative motion is very conspicuous when there is only few relative displacement. It is intuitively clear that it will be easy to find a moving stimulus among stationary distractors [4]. To model such property, a linear quantification of $\|\overrightarrow{V}_{relative}\|$ is achieved in order to build a histogram. The median value of the histogram, called Γ_{median}, is a reliable estimator of the relative motion amount. $\|\overrightarrow{V}_{relative}\|$ is then weighted by Γ_{median}. The closer Γ_{median} to 0, the

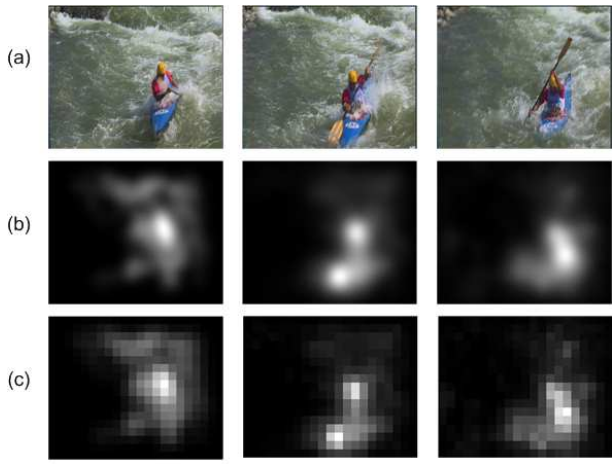

Fig. 2. Saliency maps stemming from the original pictures of the sequence *Kayak* (a); the pixel-based saliency maps (b); the macroblock-based saliency maps (c).

more the relative motion is perceptually important. Finally, the normalized temporal saliency map S^T is computed by:

$$S^T(s) = \frac{\overrightarrow{V}_{relative}(s)}{\Gamma_{median} \times \lambda},\ \lambda = \max_s(\frac{\overrightarrow{V}_{relative}(s)}{\Gamma_{median}}) \tag{2}$$

A combining process is then used [15] in order to built the final saliency map S from the spatial and the temporal map.

3 Selective H.264 video coding

3.1 Definition

In the classical approaches of video compression, no a-priori information is given on the content of the scene to be coded. At the opposite, a selective compression scheme uses information regarding localization and saliency of the interesting parts of the scene. From these information, known prior to the onset of the coding process, the coding resources, and more especially the bit budget, are distributed non-uniformly across the picture.
In this study, we use a computational model of visual selective attention to predict where viewers are likely to fixate. Such model provides a pixel-based saliency map. A simple transformation is used to transform a pixel-based saliency map into a macroblock-based saliency map. The transformation consists in averaging the saliency values located in a macroblock. Figure 2 shows examples of pixel-based and macroblock-based saliency maps.
As the global transmission rate is fixed, the encoding strategy consists in cod-

ing at coarser rate the non-relevant areas of the pictures whereas the relevant areas are encoded to a higher quality. The next section presents the proposed selective compression algorithm.

3.2 Algorithm

The proposed algorithm is a two-pass based coding framework driven by a saliency map. Before delving into the details of the algorithm, several hypothesis have to be done.
First, we assume that the bit budget B to encode a particular picture is given by a rate-control algorithm and that the quantization parameter, denoted as QP, is also known. QP is used to quantize all the macroblock of the considered picture. Second, the rate-distortion curve is supposed to be available for all macroblocks. Consequently, the encoding cost and the distortion, $c_j(i)$ and $d_j(i)$, respectively, are known for each macroblock i and for each quantization parameter j. In practice, there are several possibilities to obtain the rate-distortion curves for each macroblock [11] (empirical, statistical approaches...).
The aim is to determine for each macroblock i the best quantization parameter j around the QP value bearing in mind two key points:

(1) the first point concerns the global perceived video quality. For a given transmission rate, the global perceived video quality will be maximal if and only if the local perceived quality is the same everywhere. Therefore, the first step of the algorithm has to establish an homogeneous perceived quality across the whole picture. In others words, severe coding artifacts that stand out against its direct neighborhoods have to be avoided;
(2) the second point refers to the distribution of the remaining bit budget over the conspicuous or interesting locations in the visual scene.

A two-pass based coding framework is proposed to tackle these two important points. Each pass is described hereafter.

3.2.1 First pass: toward an uniform quality

We have previously assumed that the transmission rate is known, leading to the bit budget B. Let B^1 the total number of bits allocated for encoding the current picture during the first pass. The value B^1, strictly inferior of the bit budget B is calculated as follow: $B^1 = B \times \gamma$. The quantization parameter, denoted as QP^1, is used to reach the bit budget B^1 ($QP^1 > QP$).
The value γ depends on the target application. For example, in the case of video conference application, γ should have a low value since a strong degradation of the scene background can be tolerated. In a context of video broad-

casting, the problem is quite different because a trade-off has to be considered. It consists both in avoiding the apparition of severe artifacts over the whole picture and in having an substantial bit budget difference $B - B^1$ to increase the quality of the RoI during the second pass. As this study concerns the video broadcasting applications, γ is arbitrary set to 0.8.
What we are mainly concerned with here is to adjust locally the quantization parameters QP_i^1 associated to each macroblock i in order to obtain a maximal global perceived video quality. After the adjustment, the encoding cost of the whole picture is recalculated. The modulation of the initial quantization parameters QP^1 is based on an iterative approach consisting of three steps. Before describing in details these steps, we have to point out that the quality is measured by the mean squared errors (MSE). The MSE and the deduced peak signal-to-noise ratio (PSNR) are the most popular difference metrics in image and video processing. The MSE is the mean of the squared differences between the original and the degraded pictures. This metric is fast and easy to use. Obviously, this metric presents numerous drawbacks [6]. Nevertheless, there is currently no relevant alternative.
The first pass of the proposed algorithm is composed by the following steps:

(1) the variance or the dispersion σ^2_{MSE} of the quality values is computed over the picture;
(2) the aim is to reduce the dispersion of the objective quality in order to tend toward an uniform quality. The iterative algorithm consists in decreasing (respectively increasing) the quantization step of the M macroblocks presenting the highest distortion (respectively the lowest distortion). The bit budget B^1 is modified to account for the modification of the macroblocks encoding costs;
(3) the new dispersion σ^2_{MSE} is computed. If the difference between this value and the previous dispersion is below a threshold ϵ, the iterative process is ended, else the algorithm returns to the step 2.

In practice, M represents 10% of the total number of macroblocks. The number of iterations required to find the best adjustement depends on the ϵ value. Nevertheless, the average iteration number is close to 4.

3.2.2 Second pass: allocating the remaining bit-rate in function of the saliency map

In the literature, most of the approaches rely on a rate-distortion optimization to favor the quality of the relevant perceptual areas. A relaxation factor is included in the Lagrangian optimization, relaxing the bit rate constraint on the RoI. The Lagrangian method only finds the convex approximation, while a direct exhaustive search of constrained problem results in an optimal solution.

The exhaustive search is time consuming and it is unreasonable to use this method over the whole picture.
Nevertheless, as the size and the number of the RoI is relatively small, it is possible to conduct a direct search only on the visually important area. The process to distribute the quantity $B-B^1$ consists in seeking the macroblock for which the distortion is significantly decreased when the quantization parameter is decreased of one unit. In others words, we compute for each macroblock i, the slope of the rate-distortion curve:

$$\lambda_{j \to j-1}(i) = \frac{d_{j-1}(i) - d_j(i)}{c_j(i) - c_{j-1}(i)} \times S(i) \tag{3}$$

where, $c_x(i)$ and $d_x(i)$ represent respectively the encoding cost and the distortion for a macroblock i and a quantization parameter x. $S(i)$ represents the saliency value of the macroblock i. In order to accelerate the process, the saliency map has been thresholded. All the saliency values below a predefined threshold are set to zero. In practice, if the saliency map ranges from 0 to 255, the threshold is set to 100.
The most interesting macroblock in terms of rate-distortion is the macroblock having the highest $\lambda_{j \to j-1}$ value:

$$\lambda_{max} = \max_{i,j} \lambda_{j \to j-1}(i) \tag{4}$$

It is noticeable that a macroblock having a saliency value equal to zero will never be selected. The quantization parameter of the macroblock having the highest $\lambda_{j \to j-1}$ is decreased. The remaining bit budget $(B - B^1)$ is then updated. While the value $B - B^1$ is positive, this algorithm is repeated. It is interesting to note that a macroblock of interest can be chosen several times.

4 Performance assessment and discussion

4.1 Performance assessment

The goal of the performance assessment is to examine that the proposed approach meets the two aforementioned requirements. The first analysis concerns the quality improvement of the RoI whereas the second refers to the global perceived quality of the video sequence.
First and foremost, we have to check that the proposed selective compression scheme has the capacity to heighten the objective quality on the RoI. Table 1 shows the results in term of PSNR on three sequences (sequences

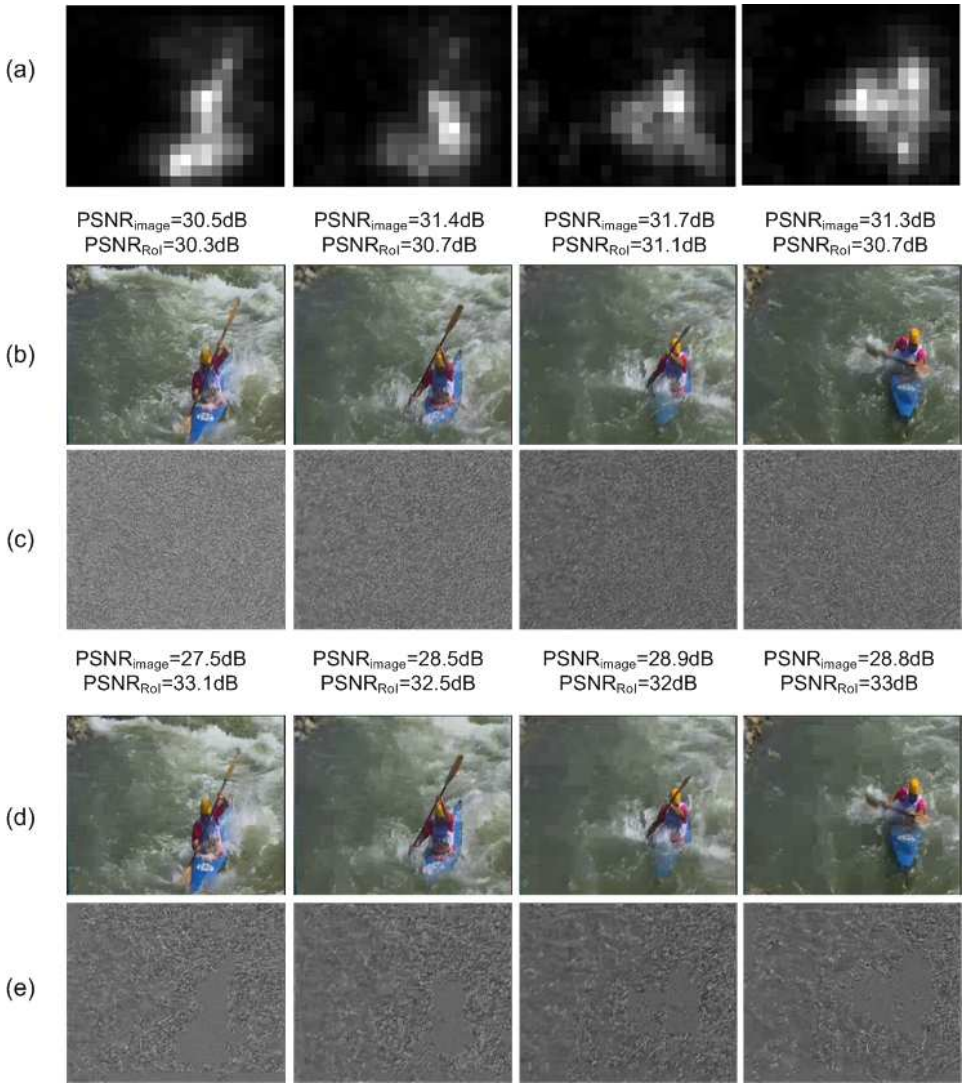

Fig. 3. Examples of selective compression on the sequence *Kayak* (CIF): (a) the macroblock-level saliency maps; (b) pictures encoded by a classical H.264 coder; (c) MSE stemming from the difference between the original and the degraded picture; (d) pictures encoded by the proposed selective compression scheme; (e) MSE stemming from the difference between the original and the degraded picture (RoI).

Kayak and *Stefan* are in CIF format whereas the last sequence has a resolution of 640×352 pixels). These results are given for several bit rates. As expected, the average PSNR computed over the whole picture coming from the classical coding is better than the average PSNR obtained by the proposed approach. However, a significant improvement of the RoI quality is achieved

Table 1
Average PSNR for three sequences for different bit rates. The goal is to compare the average PSNR computed over the whole image and over the regions of interest for a classical compression scheme and the proposed approach.

Sequence	Parameters	$PSNR_{image}$		$PSNR_{RoI}$	
		Classical coding	RoI coding	Classical coding	RoI coding
Kayak	1Mb/s, CIF	30.95	28.1	30.46	32.72
	550 Kb/s, CIF	28.87	27.59	28.66	30.13
Stefan	1Mb/s, CIF	35.71	33.43	34.52	37.45
	550 Kb/s, CIF	32.02	30.41	30.65	31.63
Raid	1.4Mb/s, 640 × 352	34.99	34.46	34.4	35.23

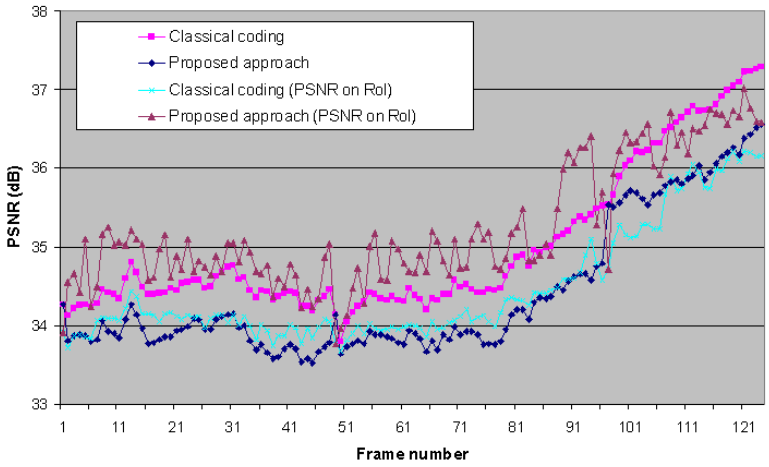

Fig. 4. Temporal evolution of the PSNR. Classical and proposed approaches are compared in term of PSNR when the whole picture and the RoI are considered.

by the proposed bit allocation strategy, regardless of the bit rate. The average improvement is about 2 dB for the CIF sequences. Concerning the sequence *Raid*, the improvement is of 0.77 dB.

Figure 4 presents the temporal evolution of the average PSNR values (PSNR computed on the whole image, on the RoI and for the classical and the proposed approach) on the sequence *Raid*. This figure shows that the difference in terms of PSNR values between the two coding schemes remains constant.

Figure 3 shows four frames of the sequence *Kayak*. The macroblock-level saliency maps are illustrated on the first row (a). Pictures encoded by the classical coding are shown on the second row (b). The MSE (figure 3 (c)), is the difference between the source and the picture encoding by the classical approach. It is noticeable that the distribution of the coding errors is uniform. The distribution of the MSE coming from the proposed approach (figure 3 (e)) is quite different. The proposed approach has drastically reduced the coding errors on the visually important areas.

Fig. 5. Examples of selective compression on the sequence *Stefan* (CIF): (a) on the left hand side, the coded picture is obtained by a classical H.264 video coding whereas, on the right hand side, the picture is obtained by the proposed approach; (b) MSE respectively associated to the classical and proposed video coding.

Another example is given on the figure 5. The quality improvement is noticeable on the face and legs of the tennis player whereas the degradation of the background is not noticeable. The sequence *Stefan* is typically the sequence on which the perceived video quality can be drastically improved by a video compression scheme driven by a saliency map: this sequence consists of a small RoI and a high textured background. Numerous bit are wasted to encode the background to the detriment of the tennis player. Figure 6 illustrates for several pictures the distribution and the evolution of the encoding cost when a classical and the proposed compression scheme are considered. It turns out that a significant part of the bit budget is wasted on visually irrelevant areas in a classical scheme (figure 6 (b)). Once again, the picture *Stefan* is a typical case where the lack of intelligence of the classical approach is obvious. Indeed, the most important part of the bit budget is used to encode the textured background, despite the fact that these areas have the capability to mask the quantization noise. The proposed approach allows to concentrate the major part of the bit budget on the visually important areas, as depicted by the figure 6 (c).

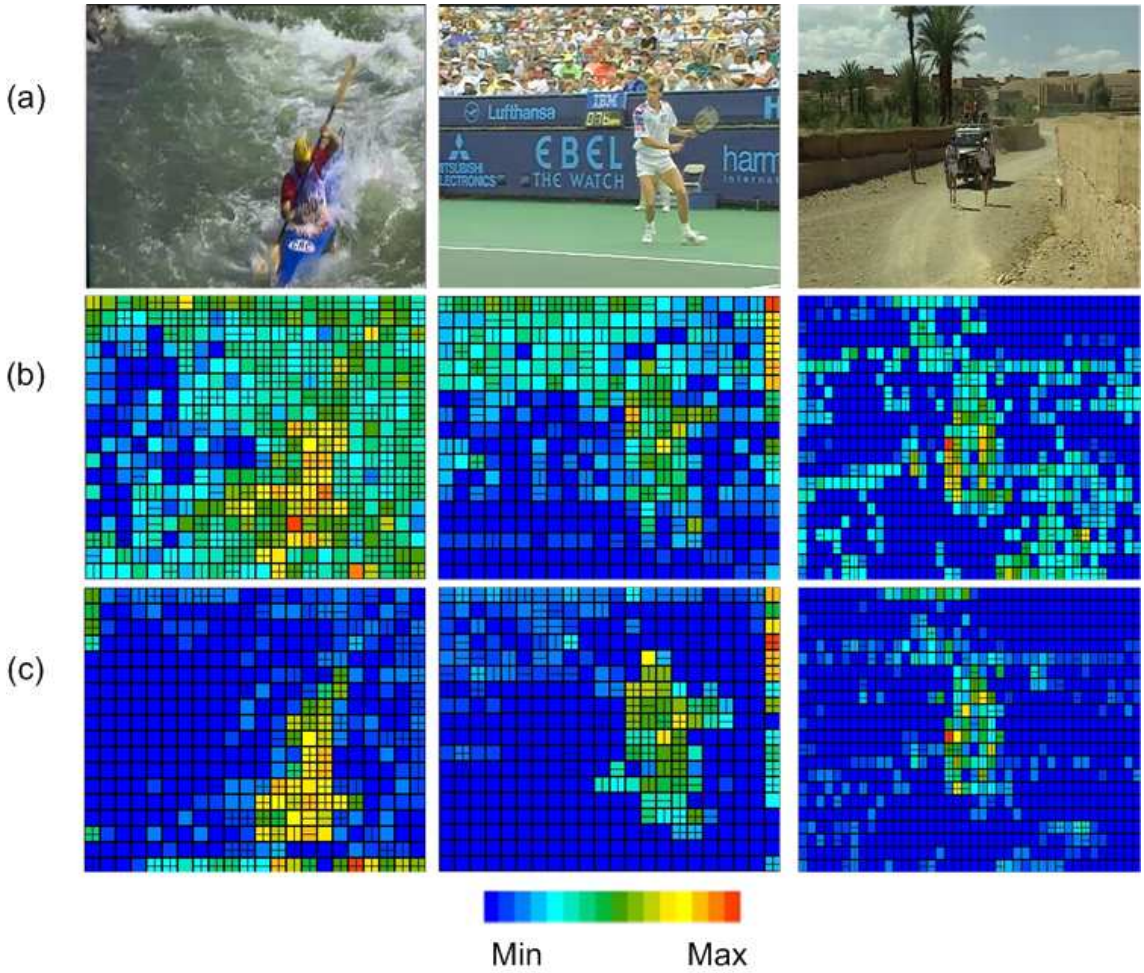

Fig. 6. Distribution of the encoding cost for a conventional H.264 coding (b) and for the proposed approach (c).

4.2 Discussion

The next future generation of coding schemes will likely use an a-priori knowledge automatically extracted from the original picture. The goal of such coding is to obtain a high fidelity reconstruction over the visually important areas in order to improve the overall perceived video quality. As Bradley mentioned in [1], the selective compression is very relevant when the size of the RoI are small and if the target bit rate is low enough to produce visible compression artifacts on the RoI. Indeed, it is clear that coding artifacts are more annoying on the RoI than in any others areas.
It is true that the aforementioned parameters act upon the relevance of the selective compression. But, there is an other parameter which is more important than the previous. It concerns the computation of the distortion caused by the coding. In the proposed approach, the simplest distortion metric (MSE) is used. This metric presents numerous drawbacks and is not convenient to control a coding strategy. A suitable metric should take into account several

important points: first and foremost, the visual masking has to be considered. Masking effect refers to the modification of the visibility of a stimulus (coding artifact, for example) due to the influences of the context, called the masking signal. Indeed, it is well known that coding artifact is more annoying on the uniform areas than on highly textured areas. Second, the distortion metric has to include the temporal dimension in order to detect flickering. Flickering annoys the viewers and can potentially attracts the visual attention, creating a new fixation points. Finally, the metric has to be dedicated to a particular compression scheme. For example, concerning the H.264 compression scheme, the metric should detect the block effect.
As a selective compression scheme using a basic distortion metric can improve the perceived quality (figure 5), the performances can further increased by the used of a reliable metric.
The assessment of the final quality is also a problem. In this study, the PSNR is used to assess the quality. This metric is widely criticized because the correlation with perceived quality measurement is small. Nevertheless, the trend obtained by the PSNR is informative and provides some insights on the interest of a selective compression scheme. A more coherent validation involving subjective quality tests will be required in order to definitively prove the interest of selective compression.

5 Conclusion

The recent breakthrough in the visual attention modeling, initiated by L. Itti [7] and pursued by [19,16,2] and by the works [14,15], allows to serenely tackle the compression scheme driven by an automatic detection of the visually important areas. The goal is to encode the region of interest at higher bit rate in order to improve the perceived visual quality. In this paper, we proposed a selective coding framework based on the recent H.264/MPEG-4 AVC video coding standard. The results are very promising. A significant improvement of the RoI, in term of PSNR, is achieved.
Besides these preliminary results, two major problems have been raised in this paper. First, in order to avoid the introduction of artifact on the background, a reliable distortion metric has to be defined taking into account both the properties of the HVS and the compression technology used to perform the encoding. The second point refers to the performance assessment of the results coming from a selective compression scheme. Classical assessment methods, based on the MSE, are not suitable anymore: a selective compression introduces more degradation in the background, increasing the average MSE. Therefore, the way to qualitatively assess the image quality has to be modified and should depend on the spatial localization of the RoI. Future development will consist in designing a new perceptual image quality assessment

method. This metric will be used to evaluate the overall perceived quality and to optimize the performances of the bit budget allocation.

References

[1] A.P. Bradley. Can region of interest coding improve overall perceived image quality? In *Proceedings of APRS Workshop on Digital Image Computing*, 2003.

[2] N. Bruce and E. Jernigan. Evolutionary design of context-free attentional operators. In *ICIP 03*, 2003.

[3] S. Daly. The visible differences predictor : An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, chapter 14, pages 179–206. MIT Press, 1993.

[4] J.H. Elder and R.M. Glodberg. Visual search asymmetries in motion and optic flow fields. *Percep. and Psycho.*, 63, 2001.

[5] J.H. Elder and R.M. Glodberg. Ecological statistics for the gestalt laws of perceptual organization of contours. *Journal of Vision*, 2, 2002.

[6] B. Girod. What's wrong with mean squared error? in A. B. Watson (ed.), MIT Press, Visual Factors of Electronic Image Communications, 1993.

[7] L. Itti, C. Koch, and E. Niebur. Model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[8] M. K. Kapadia, M. Ito, C. D. Gilbert, and G. Westheimer. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in v1 of alert monkeys. *Neuron*, 15(4):843–856, 1995.

[9] P. LeCallet. *Critres objectifs avec rfrence de qualit visuelle des images couleur.* PhD thesis, Ecole Polytechnique de l'Universit de Nantes, Universit de Nantes, 2001.

[10] A. Leontaris and P. C. Cosman. Region-of-interest video compression with a composite and a long-term frame. In *Proceedings of the Seventh IASTED International Conference Computer Graphics and Imaging*, Hawa, USA, 2004.

[11] L.J. Lin. *Video bit-rate control with spline approximated rate-distortion characteristics.* PhD thesis, University of Southern California, 1997.

[12] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004.

[13] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. Masking effect in visual attention modeling. In *Proc. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Lisbonna, Portugal, 2004.

[14] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. Performance assessment of a visual attention system entirely based on a human vision modeling. In *Proceedings ICIP-04 (IEEE International Conference on Image Processing)*, Singapor, 2004.

[15] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A spatio-temporal model of the selective human visual attention. In *Proceedings ICIP-05 (IEEE International Conference on Image Processing)*, Genoa, Italia, 2005.

[16] D.J. Parkhurst. *Selective attention in natural vision: using computational models to quantify stimulus-driven attentional allocation*. PhD thesis, The Johns Hopkins University, Baltimore, Maryland, USA, 2002.

[17] H. Snane. *Reprsentation d'images en sous-bandes visuelles. Application au codage d'images de tlvision sans dfaut visuel*. PhD thesis, IRESTE, Universit de Nantes, 1996.

[18] Y. Yang and S. S. Hemami. Rate-distortion optimizations for region and object based wavelet video coding. In *in Proc. 34th Asilomar Conference on Signals, Systems, and Computers*, 2000.

[19] H. Yee and S. Pattanaik. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. In *IACM*, 2001.