In Proceedings of IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP'04.

# PERCEPTUAL PREFILTERING FOR VIDEO CODING

*Andrea Cavallaro*

Multimedia and Vision Laboratory
Queen Mary, University of London
London E1 4NS, United Kingdom

*Olivier Steiger, Touradj Ebrahimi*

Signal Processing Institute
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

## ABSTRACT

Semantic segmentation is generally associated with second generation video coders, or object-based coders. Object-based coders encode different video objects separately in order to achieve lower bitrates and to enable object-based functionalities. In this paper, we present an encoding framework that uses semantic segmentation to improve the performance of first generation video coders, or frame-based coders. Semantic segmentation is exploited in a prefiltering step prior to encoding. This prefiltering step mimics the way humans treat visual information by separating relevant information from contextual information. Contextual information is then simplified, thus reducing the information to be coded. Experimental results on indoor as well as on outdoor test sequences when the semantics is defined by motion show that the proposed prefiltering improves the perceived quality with respect to traditional video coders.

## 1. INTRODUCTION

The increasing adoption of wireless devices is helping the development of applications requiring low bitrates, such as broadcasting sports events and news to mobile appliances, video telephony, and wireless surveillance. The success of these applications depends on user satisfaction, which in turn depends on the perceived quality of the video content delivered. In order to maximize the perceived video quality, an increasing research effort is aimed at improving video coders by taking into account human factors [1, 4, 7].

Traditional frame-based video coders treat the entire scene uniformly, assuming that people may look at every pixel of the video. In reality, humans and primates do not scan a scene in raster fashion. Our visual attention tends to jump from one point to another. These jumps are called *saccades*. Yarbus [2] demonstrated that the saccadic patterns depend on the visual scene as well as on the cognitive task to be performed. The studies of Bajcsy [3] also led to the conclusion that *we do not see, we look*. We focus our visual attention according to the task at hand and to the scene content.

Low-level as well as high-level aspects drive our visual attention. Low-level aspects include contrast, spatio-temporal frequency, color variation, texture energy, and brightness [4]. High-level aspects are related to object detection and tracking. In this paper we concentrate on the high-level aspects and in particular on semantic objects. We attempt to emulate the human visual system to prioritize the visual data in order to improve the performance of frame-based coders. Frame-based coders are widely used and there is an interest in improving their performance by adding a prefiltering step. This prefiltering step is based on semantic segmentation. The flow diagram of the proposed framework is depicted in Figure 1.

The paper is organized as follows. Section 2 describes the perceptual prefiltering step preceding the encoder. In Section 3 we introduce the methodology used for performance evaluation and in Section 4 we present the results of the proposed encoding framework. Finally, Section 5 concludes the paper and discusses future work.

## 2. PERCEPTUAL PREFILTERING

The prefiltering step aims at mimicking the way humans treat visual information. First relevant information is separated from contextual information. Then the contextual information is simplified prior to coding. We obtain the separation by decomposing each frame of the sequence to be encoded into mutually exclusive and jointly exhaustive *classes of interest*. An example is the separation of the video into two classes of interest, namely foreground and background. The definition of the semantic partition depends on the task to be performed. Therefore, some *a priori* knowledge of the objects we want to segment is required. For applications such as video conference or news broadcasting, faces may represent the semantic objects to be considered, whereas in applications such as video surveillance and sport broadcasting, motion information can be used as semantics for segmenting moving objects [5]. In the latter case, the motion of a moving object is usually different from the motion of background and other objects.

The extraction method we use to segment moving object is a change detector organized in two stages. Each stage compensates for a source of noise, namely camera noise and local illumination variations. Camera noise is reduced in a classification stage which takes into account the statistics of the input video and adapts the detection threshold to local information [5]. The second stage reduces local illumination variations and produces a spatio-temporal regularization of the classification results [6]. Figure 2(b) shows and example of segmentation corresponding to the results presented in Section 4.

This decomposition of the scene into meaningful objects prior to encoding is used in the perceptual prefiltering. The areas belonging to the foreground class, or semantic objects, are used as region of interest. The areas not included in the region of interest may either be eliminated, that is set to a constant value, or lowered in importance by using a low-pass filter. The latter solution simplifies the information in the background, while still retaining essential contextual information.

The above-mentioned prefiltering step can be used without a specific knowledge of the encoder. In case there is access to the

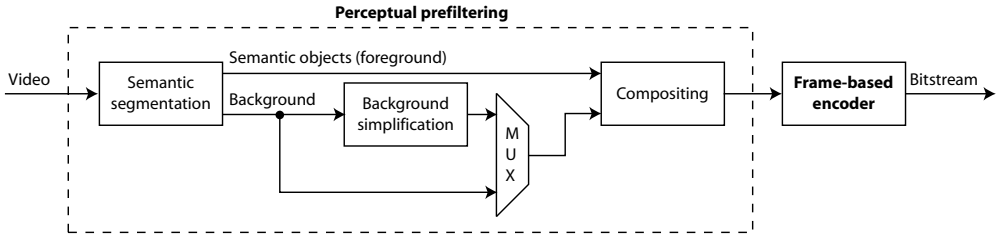


**Fig. 1**. The proposed encoding framework based on perceptual prefiltering.

(a) (b)

**Fig. 2**. Example of semantic segmentation: (a) original frame; (b) semantic objects.

specific coder being used (e.g., MPEG–1), another way to lower the importance of less relevant portions of an image before coding is to take advantage of the characteristics of the coding algorithm. In the case of block-based coding, each background macroblock can be replaced by its DC value. The final effect is equivalent to that described previously, but optimized for the specific encoder.

## 3. PERFORMANCE EVALUATION

A combination of subjective and objective evaluation techniques is used to assess the performance of a coder with perceptual prefiltering. Subjective evaluation includes the visual comparison of frames and frame details. Objective evaluation includes temporal signal-to-noise ratio analysis. To account for the way humans perceive visual information, different parts of an image, or object classes, should be considered [7, 8]. As opposed to traditional MSE, object classes are taken into account through a distortion measure, here referred to as the *semantic mean squared error*, SMSE, defined as:

$$\mathrm{SMSE} = \sum_{k=1}^{N} w_k \cdot \mathrm{MSE}_k, \tag{1}$$

where $N$ is the number of object classes and $w_k$ the weight of class $k$. Class weights are chosen depending on the semantics, with $w_k \geq 0, \forall k = 1, \ldots, N$ and $\sum_{i=1}^{N} w_k = 1$. The mean squared error of each class, $\mathrm{MSE}_k$, can be written as

$$\mathrm{MSE}_k = \frac{1}{|C_k|} \sum_{(i,j) \in C_k} d^2(i,j), \tag{2}$$

where $C_k$ is the set of pixels belonging to the object class $k$ and $|C_k|$ is its cardinality. The class membership of each pixel $(i, j)$ is defined by semantic segmentation. The error $d(i, j)$ between the original image $I_O$ and the distorted image $I_D$ in Eq.(2) is the pixel-wise color distance. The color distance is computed in the 1976 CIE $Lab$ color space in order to consider perceptually uniform color distances with the Euclidean norm and is expressed as:

$$d(i,j) = \sqrt{\left(\Delta I^L(i,j)\right)^2 + \left(\Delta I^a(i,j)\right)^2 + \left(\Delta I^b(i,j)\right)^2}, \tag{3}$$

with $\Delta I^L(i,j) = I_O^L(i,j) - I_D^L(i,j)$, $\Delta I^a(i,j) = I_O^a(i,j) - I_D^a(i,j)$, and $\Delta I^b(i,j) = I_O^b(i,j) - I_D^b(i,j)$. The final quality evaluation metric, the *semantic peak signal-to-noise ratio*, SPSNR, is the following:

$$\mathrm{SPSNR} = 10 \log_{10} \left( \frac{V_{\max}^2}{\mathrm{SMSE}} \right), \tag{4}$$

where $V_{\max}$ is the maximum peak-to-peak value of the color range. When the object classes are foreground and background, then $N = 2$ in Eq.(1). If we denote with $w_f$ the foreground weight, then $\mathrm{SPSNR} \equiv \mathrm{PSNR}$ when $w_f = 0.5$. The larger $w_f$, the more important the contribution of the foreground. When $w_f = 1$, then only the foreground is considered in the evaluation of the peak signal-to-noise ratio.

## 4. RESULTS

In this section, we present the evaluation of the proposed encoding framework based on perceptual prefiltering. Sample results from the test sequences *Hall monitor*, from the MPEG–4 video content set, and *Highway*, from the MPEG–7 video content set, are shown. Both sequences are in CIF format at 25 Hz. The sequences have been coded in MPEG–1 *with* and *without* perceptual prefiltering using TMPGEnc 2.521.58.169 software with constant bitrate (CBR) rate control. Five modalities have been considered in the comparison: (1) original sequence; (2) temporal resolution reduction (from 25 frames/s to 12.5 frames/s); (3) spatial resolution reduction (from CIF to QCIF); (4) video objects composited with static background; (5) video objects composited with lowpass-filtered background. Modalities (1) to (3) do not use perceptual prefiltering, whereas modalities (4) and (5) use perceptual prefiltering.

Figure 3 shows the SPSNR ($w_f = 0.8$) of the test sequences coded at 150 Kbit/s with the modalities (1), (4) and (5). The average improvement obtained by perceptual prefiltering with static background (4) over the original sequence (1) is of 2.1dB for *Hall monitor* and 1.41dB for *Highway*. The average improvement obtained by perceptual prefiltering with lowpass-filtered background

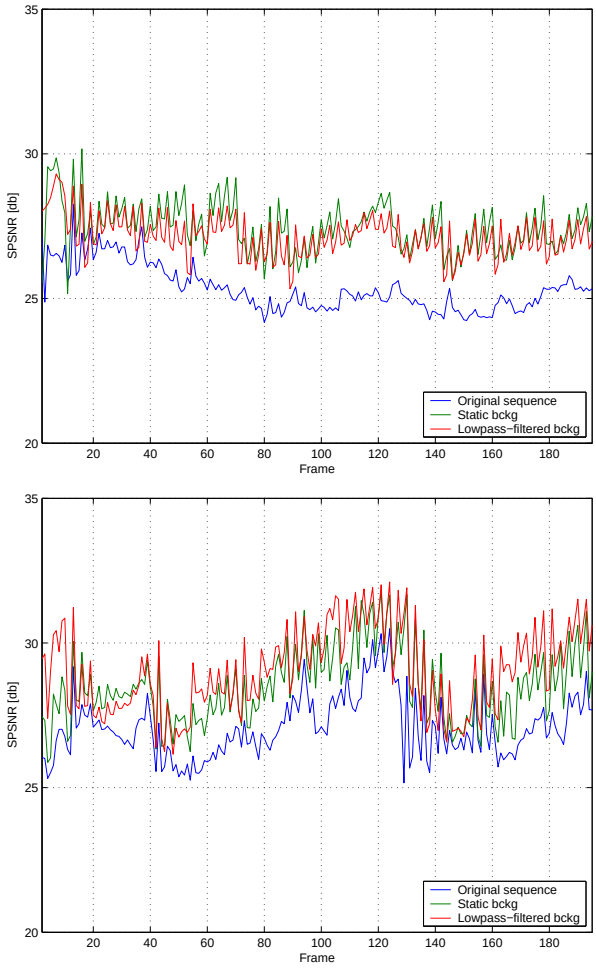


**Fig. 3**. The SPSNR of (top) *Hall monitor*; (bottom) *Highway* coded with MPEG–1 at 150 Kbit/s *with* and *without* perceptual prefiltering.

(5) over the original sequence (1) is of 1.78dB for *Hall monitor* and 1.96dB for *Highway*. Note that the measured quality of the semantic modalities (4) and (5) is almost always larger than that of (1). The few cases in which the performance is not better correspond to frames where none or a very limited portion of the picture is occupied by the foreground: in the sequence *Hall monitor* this happens before the two persons enter the corridor, in the sequence *Highway* when there are only two cars far away from the camera. In these cases, the perceptually filtered background, which differs from the original background, is responsible for a lower overall quality. This phenomenon is not annoying when the sequence is seen as a whole. The values of SPSNR measured for all modalities (1) to (5) are summarized in Table 1.

Figure 4 shows a sample frame from each test sequence coded with MPEG–1 at 150 Kbit/s with and without perceptual pre-filtering. The modalities used in this subjective comparison are the following: (a) original sequence; (b) static background; (c) lowpass-filtered background. It is possible to notice that the objects are better reconstructed when perceptual prefiltering is used. Figure

| SEQUENCE | MODALITY | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| *Hall monitor* | 25.22 | 26.38 | 22.12 | 27.32 | 27.00 |
| *Highway* | 26.95 | 24.61 | 22.22 | 28.36 | 28.92 |

**Table 1**. Average SPSNR (in dB) with modalities (1) to (5) for the test sequences *Hall monitor* and *Highway* coded with MPEG–1 at 150 Kbit/s.

5 shows magnified excepts from both test sequences. The first row shows the person carrying a monitor in *Hall monitor*. The amount of coding artifacts is notably reduced by perceptual prefiltering (b,c). In particular, the person's mouth and the monitor are visible in (c), whereas they are corrupted by coding artifacts in the non-semantic modality. Similar observations can be made for the second row of Figure 5, which shows a blue truck entering the scene at the beginning of the *Highway* sequence. Coding artifacts are less disturbing on the object in (b) and (c) than in (a). Moreover, the front-left wheel of the truck is only visible with perceptual prefiltering.

(a) (b) (c)

**Fig. 5**. Details of frame 280 of *Hall monitor* (top) and frame 16 of *Highway* (bottom) using different modalities: (a) original sequence; (b) static background; (c) lowpass-filtered background.

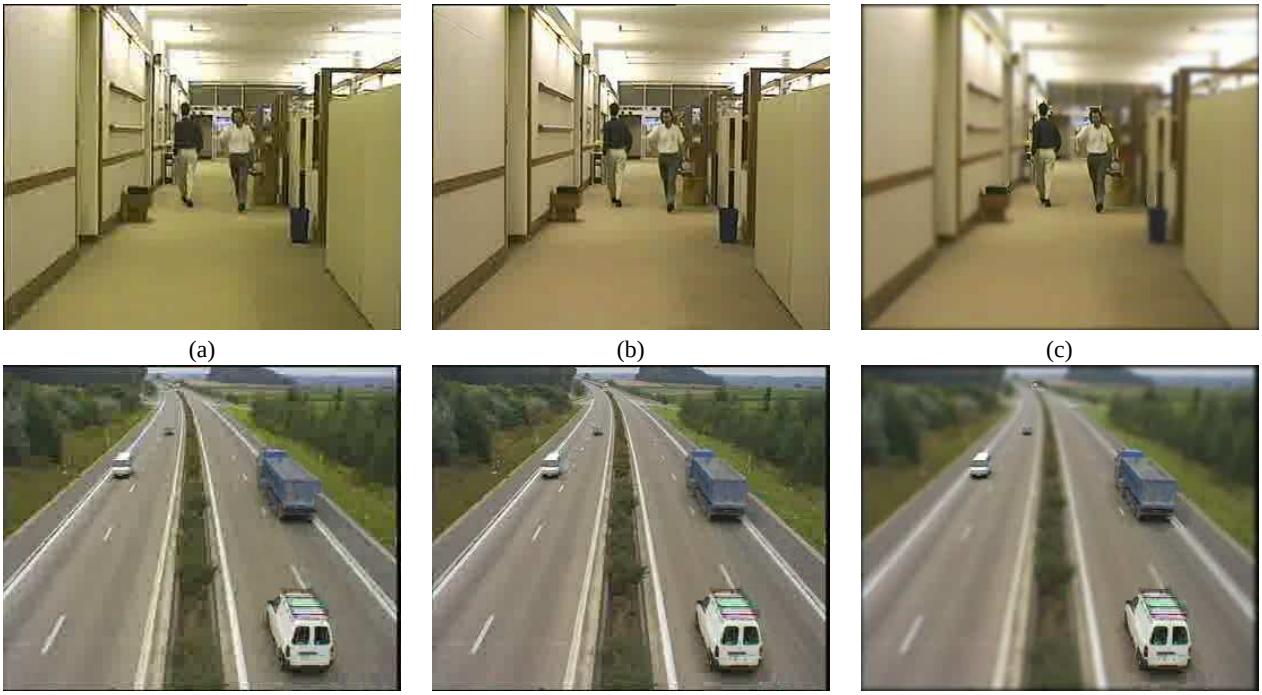

**Fig. 4**. Frame 190 of *Hall monitor* (top) and frame 44 of *Highway* (bottom) coded with MPEG–1 at 150 Kbit/s using different modalities: (a) original sequence; (b) static background; (c) lowpass-filtered background.

## 5. CONCLUSIONS

We presented a method for improving the performance of frame-based coders which is based on the use of semantics for prefiltering. In particular, the effectiveness of the proposed method in improving the perceptual quality at low bitrates has been demonstrated in sequences containing moving objects. The prefiltering step is general, can be used with other types of video objects, such as faces, and does not require to be aware of the particular frame-based coder used.

Future work includes three main research directions: (i) the integration of lower level aspects of vision in the prefiltering step; (ii) extensive performance evaluation based on subjective testing; (iii) the optimization of the prefiltering step for specific encoders with the introduction of a feedback loop.

## 6. REFERENCES


[1] Z. Lu, W. Lin, X.K. Yang, E.P. Ong, S.S. Yao "Spatial selectivity modulated Just-Noticeable-Distortion profile for video," in *Proc. Int. Conf. on Acoustic, Speech, and Signal Processing*, Montreal, Canada, vol. III, pp. 705–708, 2004.

[2] A. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.

[3] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 996–1005, 1988.

[4] Anthony Maeder, Joachim Diederich, and Ernst Niebur, "Limiting Human Perception for Image Sequences," in *Proceedings of Conference on Human Vision and Electronic Imaging*, San Jose, CA, vol. 2657, pp. 330–337, 1996.

[5] A. Cavallaro, T. Ebrahimi, "Accurate video object segmentation through change detection," in *Proc. Int. Conf. on Multimedia and Expo*, Lausanne, Switzerland, 2002.

[6] A. Cavallaro, E. Salvador, and T. Ebrahimi, "Shadow detection in image sequences," in *Proc. of IEE Conference on Visual Media Production*, London, UK, 2004, pp. 165–174.

[7] R. Cucchiara, C. Grana and A. Prati, "Semantic Transcoding for Live Video Server," in *Proc. of ACM Conf. on Multimedia*, pp. 223–226, Juan–Les–Pins, France, 2002.

[8] X.K. Yang, W.S. Lin, Z.K. Lu, E.P. Ong, S.S. Yao, "An effective perceptual weighting model for videophone coding," in *Proc. Int. Conf. on Acoustic, Speech, and Signal Processing*, Montreal, Canada, vol. III, pp. 145–148, 2004.