EVALUATING PERCEPTUALLY PREFILTERED VIDEO

Olivier Steiger, Touradj Ebrahimi

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute
CH-1015 Lausanne, Switzerland
{olivier.steiger,touradj.ebrahimi}@epfl.ch

Andrea Cavallaro

Multimedia and Vision Lab
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
andrea.cavallaro@elec.qmul.ac.uk

ABSTRACT

Perceptual prefiltering is the process of enhancing relevant portions of an image or of a video, and of simplifying contextual information in order to improve the perceived quality or the compression ratio. In this paper, we discuss the results of subjective quality evaluation experiments performed to assess the impact of perceptual prefiltering on video coding and we propose an objective quality metric that mimics the behavior of human observers. The predicted performance of the proposed metric is consistent with the subjective evaluation scores. Experimental results demonstrate that perceptual prefiltering leads to quality improvements by up to 10% at low bitrates.

1. INTRODUCTION

Perceptual prefiltering aims at mimicking the way humans treat visual information in order to improve the compression ratio of image and video coders. The overall image quality can be improved by degrading image areas that are not expected to attract the attention of a viewer in order to improve the quality (i.e. the associated bit allocation) of areas that observers are looking at [1]. To enable perceptual prefiltering, relevant portions of visual information (foreground) need to be separated from contextual information (background).

Previous work on perceptual prefiltering is based mainly on low-level features. Non-linear integration of low-level visual cues that mimics the processing in primate occipital and posterior parietal cortex is used in [2]. Visual cues are combined into a saliency map that modulates encoding priority. In [3], block importance is determined directly in the DCT domain by using a discontinuity height measure, which gives the contrast of dominant discontinuities within the block. Highly contrasted discontinuities are considered to be visually important. Other methods consider high-level information (semantics) [4, 5]. In [4], each frame frame is subdivided into a number of classes of relevance that are coded at a different level of quality by an object-based encoder. The definition of the classes depends on the task to be performed. For applications such as video conference or news broadcasting, faces may represent the classes to be considered, whereas in applications such as video surveillance and sport broadcasting, motion can be used for segmenting moving objects. In [5], each frame of the sequence is separated into foreground and background classes based on motion information. Then, after background simplification, both parts are re-composited together and coded by a frame-based encoder (Fig. 1).

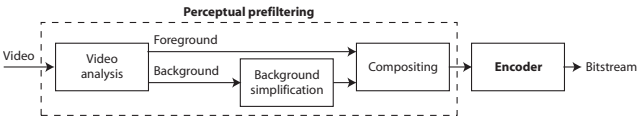


Fig. 1. Perceptual prefiltering based on semantic information is the process of video analysis, followed by background simplification and compositing.

In this paper, we quantify the impact of perceptual prefiltering with subjective experiments and we show that background alterations resulting from perceptual prefiltering do not impair overall quality at low bitrates. Moreover, we propose an objective quality metric that mimics the behavior of human observers. The metric overcomes the limitations of subjective evaluation experiments that are expensive, time consuming and cannot be used to assess video quality in real time.

The paper is organized as follows. Subjective experiments are reported and discussed in Section 2. The objective quality metric is presented and validated in Section 3. Finally, we draw conclusions in Section 4.

2. SUBJECTIVE EVALUATION

2.1. Experimental setup

Four test sequences from the MPEG–4 Video Content Set are used for subjective performance evaluation: *Children*, *Coastguard*, *Hall monitor* and *Akiyo*. The sequences include deforming and rigid objects of different size, complex as well as simple background, and different types of motion. The TMPGEnc 2.521.58.169 MPEG–1 codec with constant bitrate (CBR) rate control is used. The coding structure is 'IBBPBBPBBPBBPBBPBB'. Bitrates are chosen so as to range from the lowest bitrate supported by the codec, up to perceptually lossless coding. Since we expect results to stabilize at high bitrates, tested rates are distributed exponentially: 200, 250, 300 and 500 Kbit/s for all sequences, plus 150 Kbit/s for *Akiyo* & *Hall monitor*, and 100 Kbit/s for *Akiyo*.

Perceptual prefiltering is either achieved by lowpass filtering, or by replacing the original background by a static background shot (*Hall monitor*). The foreground is hand-segmented in order to avoid bias due to segmentation errors. The preprocessing methods under analysis are: (1) spatial resolution reduction; (2) perceptual prefiltering with lowpass filtering; (3) perceptual prefiltering with static background. In Fig. 2, a sample frames from each sequence

0-7803-9332-5/05/$20.00 ©2005 IEEE

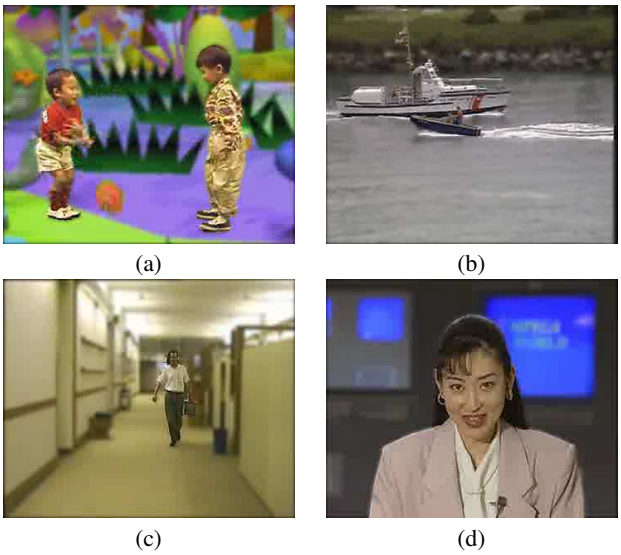


Fig. 2. Sample frame coded with MPEG–1 at 150 Kbit/s using perceptual prefiltering with lowpass filtering. (a) *Children*. (b) *Coastguard*. (c) *Hall monitor*. (d) *Akiyo*.

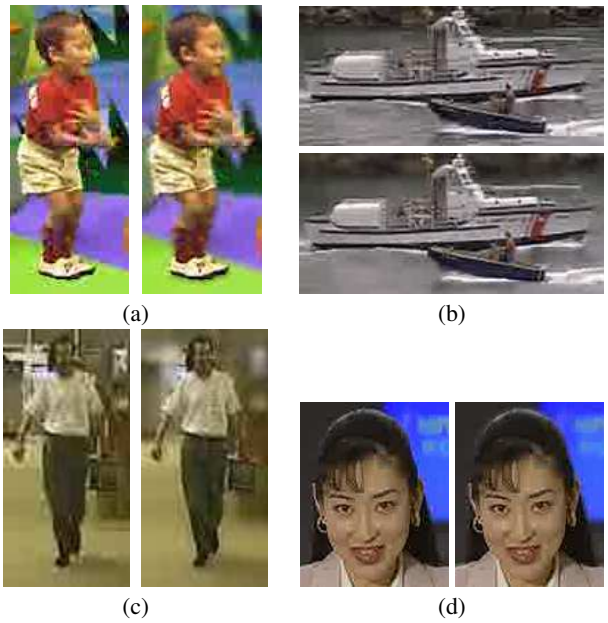


Fig. 3. Frame details with and without semantic prefiltering. (Left/top) coded original; (right/bottom) perceptual prefiltering with lowpass filtering. (a) *Children*. (b) *Coastguard*. (c) *Hall monitor*. (d) *Akiyo*.

coded with MPEG–1 at 150 Kbit/s using perceptual prefiltering with lowpass filtering is given. Clearly, lowpass filtering of the background does not inhibit the main content message.

The conditions for subjective evaluation experiments follow the *Absolute Category Rating* (ACR) evaluation method, according to ITU-T Recommendation P.910 [6]. ACR is well-suited for qualification tests (i.e., to compare the performance of different coding strategies), as the method does not use explicit references. Twenty non-expert observers of different ages and backgrounds are presented a series of video sequences in random order; the presentation order is modified for each observer. Each observer participate to one sessions and each session contains 75 presentations. After each presentation, observers rate the quality of the sequence on a scale ranging from 0 (bad) to 100 (excellent). The presentation duration is 8 seconds and maximum 10 seconds are allowed for voting. Before each session, the range of qualities is presented to the observers in a training phase.

2.2. Statistical analysis of subjective evaluation results

Subjective experiments produce distributions of integer values, each number corresponding to one vote. These distributions exhibit a number of variations due to the difference in judgement between observers, and to the effect of a variety of conditions associated with the experiment. Specifically, a *session* consists of a number of *presentations* L. A presentation is obtained by applying one of a number of *test conditions* J, to one of a number of *test sequences* K. Each combination of test sequence and test condition may be *repeated* a number of times R. The *mean score* for each presentation, $\overline{u}_{jkr}$, is then given by

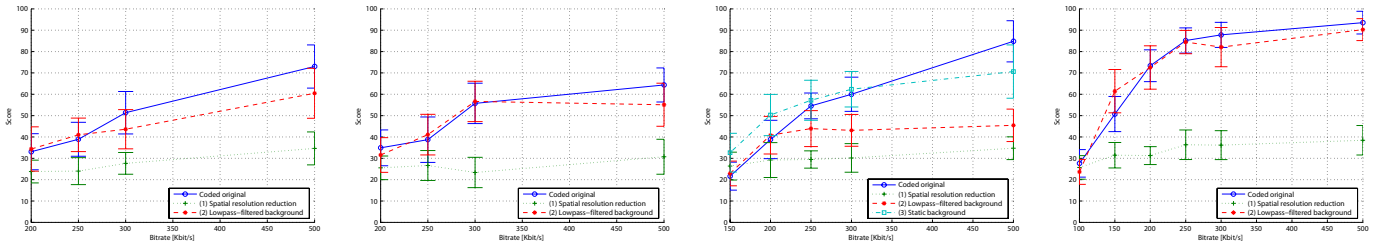
$$\overline{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr}, \qquad (1)$$

where u_{ijkr} is the score of observer i for test condition j, sequence k, and repetition r. N is the total number of observers. The associated *confidence interval* is derived from the standard deviation and size of each sample. It is proposed to use the 95% confidence interval, which is given by $[\overline{u}_{jkr} - \delta_{jkr}, \overline{u}_{jkr} + \delta_{jkr}]$, where $\delta_{jkr} = 1.96 \cdot (S_{jkr}/\sqrt{N})$. S_{jkr} is the standard deviation for each presentation.

Votes from unreliable observers are discarded aid of a *screening procedure*, organized in two stages. The first stage ensures that responses were entered accurately and in accordance with the experimental instructions. In the second stage, the variability of the data is reduced using the two-step method described in Annex 2 of ITU-R Recommendation BT.500-11 [7]. First, an expected range of values is calculated for each presentation. Then, the expected ranges are applied to the judgement of each observer. Finally, a subject is rejected for being erratic on both sides of the range, but not for being always above or always below the expected range. The results of subjective quality evaluation experiments are summarized in Figure 4. The graphs show the mean quality and associated 95% confidence interval as a function of coding bitrate.

2.3. Discussion

From Fig. 4, it is possible to notice that perceptual prefiltering (2-3) has a positive impact at low bitrates, in particular when the original background is replaced by a static frame or a sprite representing the background (3) (*Hall monitor*). At bitrates up to 300 Kbit/s, this increases the mean quality by up to 10 points as compared to the coded original. This is because inter-coded, static background blocks do not produce residue, so most of the available bitrate can be allocated to foreground objects.



(a) (b) (c) (d)

Fig. 4. Subjective evaluation results of perceptually prefiltered video. The graphs show the mean quality and associated 95% confidence interval as a function of bitrate. (a) *Children*. (b) *Coastguard*. (c) *Hall monitor*. (d) *Akiyo*.

Lowpass-filtering (2) has a lesser impact. Viewers notice the improvement of foreground quality due to the additional bandwidth freed by the filter, but at the same time they are annoyed by the loss of background information. For *Akiyo*, the quality of lowpass-filtered and coded original versions is similar over the entire bitrate range. This is because the background of the original sequence is out of focus, and thus has few high-frequency components. For *Hall monitor*, the mean quality of lowpass-filtering is slightly above that of the coded original (+1.5) at bitrates up to 200 Kbit/s. The same is true (+1.3) for *Children* at bitrates up to 250 Kbit/s. For Coastguard, lowpass-filtering has been rated above the coded original (+2.5) at bitrates of 250 and 300 Kbit/s, but below (-3.5) at the lowest bitrate of 200 Kbit/s. This is because at 200 Kbit/s, foreground objects are corrupted by heavy artifacts in both versions, whereas at 250 and 300 Kbit/s, lowpass-filtering notably reduces artifacts that are still visible in the coded original. The improvement of foreground quality can be verified in Fig. 3. Semantic prefiltering notably enhances the face in *Children* and the boats in *Coastguard*.

Background simplifications resulting from perceptual prefiltering (2-3) do not penalize overall quality at low bitrates (100-250 Kbit/s). In fact, image degradations are strong at such bitrates, and improvements on important image parts due to the additional bandwidth freed by background simplification are positively perceived. At high bitrates on the other hand, both foreground and background are coded at high quality. Thus, background alterations are easily noticed by observers and degrade the overall impression.

3. OBJECTIVE EVALUATION

3.1. Quality metric

Subjective evaluation experiments are expensive, time consuming and cannot be used to assess video quality in real time. An objective evaluation metric would therefore be desirable. An objective video distortion measure that emulates human judgement needs to account for different image areas and their relevance to the observer. This aspect can be considered with the traditional Mean Squared Error (MSE) by weighting different image areas according to their semantics. This leads to the *semantic mean squared error*, SMSE, defined:

$$\mathrm{SMSE} = \sum_{k=1}^{N} \frac{w_k}{|C_k|} \sum_{(i,j)\in C_k} d^2(i,j), \qquad (2)$$

where N is the number of 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 $\sum_{i=1}^{N} w_k = 1$. C_k is the set of pixels belonging to the object class k, and $|C_k|$ is its cardinality. 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. The final quality evaluation metric, the *semantic peak signal-to-noise ratio* SPSNR, uses SMSE instead of MSE as compared to PSNR. When the classes are foreground and background, then $N = 2$ in Eq. (2), and w_f is the foreground weight. The background weight is thus $(1 - w_f)$. The value of w_f is computed as described in the following section.

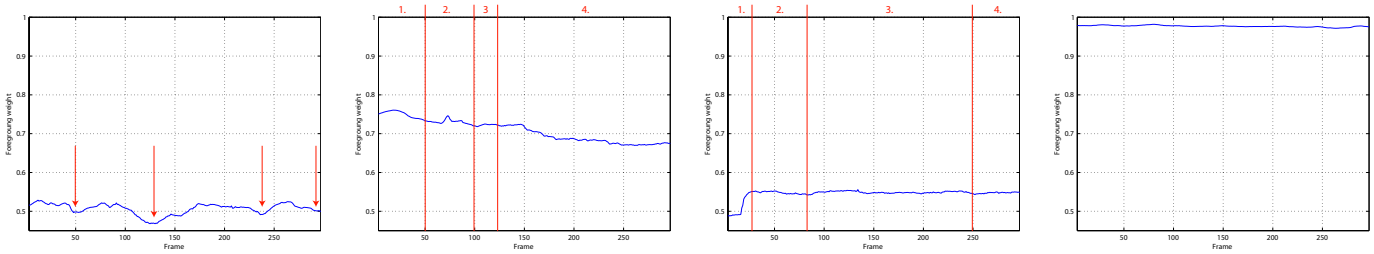
3.2. Foreground relevance

Subjective experiments quantify the amount of attention that we pay to the foreground and to the background. The foreground weight, w_f, is determined by minimizing the Pearson correlation [8] between SPSNR and subjective results. For the sequence *Akiyo*, where the foreground covers a large area and the background is simple, the observers focused mostly on foreground, thus leading to a value of $w_f = 0.97$. For *Hall monitor*, whose background is more complex and objects are smaller, the foreground attracted less attention ($w_f = 0.55$). The sequence *Children* has a very complex and colored background that attracted the observer's attention, thus resulting in foreground and background being equally weighted ($w_f = 0.5$). The sequence *Coastguard* contains camera motion. This prevented the observer from focusing on background steadily, even though it is quite complex. In this case, $w_f = 0.7$. In general, results confirm that large moving objects and complex background tend to attract users attention.

Based on the data collected with subjective experiments, we predict the foreground weight based on the following formula:

$$w_f = (\alpha - \beta \cdot \sigma_b) \cdot r + \gamma \cdot v + (\sigma_b + 1) \cdot \delta, \qquad (3)$$

where r represents the portion of the image occupied by foreground pixels: $r = |C_f|/(|C_f| + |C_b|)$, with $|C_f|$ and $|C_b|$ representing the number of foreground and background pixels, respectively. The background complexity is taken into account with σ_b, the standard deviation of the luminance of background pixels. The presence of camera motion is considered with v: $v = 1$ for moving camera, and $v = 0$ otherwise. α, β, γ and δ are constants whose values have been determined based on the results of the subjective experiments: $\alpha = 5.7$, $\beta = 0.108$, $\gamma = 0.2$ and $\delta = 0.01$.



(a) (b) (c) (d)

Fig. 5. Foreground weight, w_f, as a function of time. (a) *Children*. (b) *Coastguard*. (c) *Hall monitor*. (d) *Akiyo*.

3.3. Discussion

Eq. 3 has been used to compute the foreground weight, w_f , as a function of time. The corresponding graphs are shown in Fig. 5, where important content segments are highlighted. The results reflect the fact that observer's attention tends to be attracted by large moving objects and complex background. The foreground weight of the sequence *Children* goes through four local minima in the vicinity of frames 50, 130, 235 and 290. Each minimum corresponds to one of the children kneeling down to pick up the ball. As a consequence, the portion of the image occupied by foreground pixels, r, decreases, and the temporarily uncovered background tends to attract some additional attention. The action in *Coastguard* is not clearly perceptible in Fig. 5(b). The reason is that in Eq. 3, fluctuations of the background complexity resulting from background illumination changes due to the moving camera, σ_b, affect the foreground weight to a larger extent than variations of the image portion occupied by foreground pixels, r. For *Hall monitor*, w_f increases when r increases as well. For instance, the average foreground weight in the first segment, where the first person enters the room, is $w_f = 0.49$, whereas $w_f = 0.55$ in the third segment, where both people are visible. This reflects the fact that large moving objects tend to attract the attention. Finally, w_f is almost constant for *Akiyo*, since both σ_b and r do not show any significant variations. This is due to the fact that there is no change in the filmed action.

The prediction performance of the SPSNR metric with respect to subjective ratings is characterized by its *accuracy*, *monotonicity* and *consistency*. *Accuracy* is given by Pearson linear correlation coefficient r_p, *monotonicity* by Spearman rank-order correlation coefficient r_s, and *consistency* by outlier ratio r_o [8]. Pearson correlation, r_p, and Spearman correlation, r_s, are close to 1 for all sequences. Thus, accuracy and monotonicity of SPSNR are high. Outlier ratio, r_o, is in the vicinity of 10%, so the consistency of the metric is good as well. By comparing the Pearson correlation of SPSNR with the Pearson correlation of PSNR, we further note that by taking into account semantics, accuracy is improved by up to 8% (*Akiyo*).

4. CONCLUSIONS

The effectiveness of perceptual prefiltering in improving the quality at low bitrates has been quantified and analyzed subjectively and objectively. Moreover, an objective video distortion measure, SPSNR, that emulates human judgement has been described that accounts for different image areas and their relevance to the observer. Subjective experiments have confirmed that large moving objects and complex background tend to attract observer's attention. At low bitrates, perceptual prefiltering improves quality by up to 10%. In particular, the replacement of the background with a still background shot results in significantly more bandwidth being allocated to important image regions, without degrading the overall quality. This is very important for applications with fixed cameras, such as news broadcast and video surveillance, or when it is possible to compute a sprite of the background, such as sport broadcasting.

5. REFERENCES

[1] A. P. Bradley and F. W. M. Stentiford, "Visual attention for region of interest coding in JPEG 2000," *Journal of Visual Communication and Image Representation*, vol. 14, pp. 232–250, 2003.

[2] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[3] T. Kim and J. S. Choi, "Content-based video transcoding in compressed domain," *Signal Processing: Image Communication*, vol. 17, pp. 497–507, 2002.

[4] R. Cucchiara, C. Grana, and A. Prati, "Semantic transcoding for live video server," in *Proc. of Tenth ACM Int. Conf. on Multimedia*, 2002, pp. 223–226.

[5] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Perceptual prefiltering for video coding," in *Proc. of IEEE Int. Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP'04*, October 2004.

[6] ITU, "Subjective video quality assessment methods for multimedia applications," Tech. Rep. P.910, ITU-T Recommandation, September 1999.

[7] ITU, "Methodology for the subjective assessment of the quality of television pictures," Tech. Rep. BT.500-11, ITU-R Recommandation, 2002.

[8] D. Freedman, R. Pisani, and R. Purves, *Statistics*, W.W. Norton & Company, 3 edition, 1997.