QUALITY ASSESSMENT OF IMAGE FEATURES IN VIDEO CODING

Christian J. van den Branden Lambrecht [*] *Olivier Verscheure* [†]
Jérôme Urbain [‡] *Florent Tassin* [‡]

Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland
[†] Telecommunications Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland

ABSTRACT

This paper describes an extension of a vision model for video that has been designed to estimate how specific features in moving pictures are rendered as a consequence of a compression process. The resulting model permits to evaluate the distortions on contours and textures in a sequence, as well as to estimate how strong the blocking effect (resulting from a block-DCT compression scheme) is. The model has been used to evaluate feature rendition in MPEG-2 compressed sequences.
Keywords: Quality assessment, vision model, MPEG, test, video

1. INTRODUCTION

There has been, over the past few years, a growing interest in the field of test and measurement for digital video. The reason is that the recent developments in digital compression and communication techniques now make possible the release of a new generation of products using digital compression, such as the MPEG-1 and MPEG-2 standards.

Consequently, the question on how these new devices and systems have to be tested became important and the scientific community began to investigate possible solutions [2, 6, 8]. The present work is the continuation of the approach proposed in [9]. In that later work, a spatio-temporal vision model had been introduced and parameterized for a video compression framework. A computational quality metric had been built on the basis of this model and used to assess the quality of MPEG-2 compressed sequences. A key notion in [9] was the introduction of quality measurements on *image features*. The principle of the approach is that it is important to know how basic features in images, such as contours, textures or uniform areas are rendered in order to have better insights on a compression system's performance. The implementation that had been proposed in [9] was fairly simple, however, and this work proposes new tools to carry out the proposed task.

The paper is divided as follows: The vision model is outlined in Sec. 2 and the specific metrics for feature distortions are introduced in Sec. 3 along with computer simulations. Section 4 eventually concludes the discussion.

[*] Now with Hewlett-Packard Laboratories, Palo Alto, CA
[‡] Visiting student from Faculté Polytechnique de Mons, Belgium

2. THE VISION MODEL

The vision model used in this work is a multi-channel model for video, introduced in [7]. It takes as input an original video sequence and a distorted version of the precedent. The sequences are convolved with a set of spatio-temporal filters that emulates the collection of detection mechanisms of the visual cortex (termed *channels*). There are five spatial frequency bands (organized in octave bands), four orientation bands and two temporal channels. After the convolution, a model of pattern sensitivity is then applied to the data. This is done by computing the detection threshold (i.e the probability of seeing the distortion) for the distortion on a pixel and channel basis, accounting for contrast sensitivity of the eye and visual masking (interferences between two stimuli, i.e. between the original scene and the distortion). Further details are described in [7].

3. METRICS FOR IMAGE FEATURES

The proposed system is based on the block diagram outlined in Fig. 1. The front end vision model, identical to the one described in [9] assesses visibility of the distortion, accounting for the multi-resolution architecture of the primary visual cortex, of contrast sensitivity and visual masking. Once the amount of *perceived* distortion is known, the assessment of the visibility of distortion on image features is refined by proceeding as follows: a simple segmentation scheme is applied in order to determine which areas are textures, contours and uniform areas. The segmentation tool is the one described in [3]. Then, having a model of the distortion, we better assess how visible it is by robust spectral estimation. The three tools that are proposed are:

3.1. Contour Distortion

Contour distortion in DCT-based coding scheme is known as "mosquito noise", which gives the impression of a fake contour, perpendicular to the actual one. This effect is actually due to an ambiguity on the phase of the reconstructed DCT values. Let $B_{n,m}$ be a DCT basis function, defined over an 8×8 block, the expression of which is:

$$B_{n,m}(k,l) = K \cos\left[\frac{\pi}{16}(2n+1)k\right] \cos\left[\frac{\pi}{16}(2m+1)l\right] ,$$

where K is a constant, n, m define the index of the basis function and k, l are the coordinates within the 8×8 block. The above expression can be written, expanding the cosines

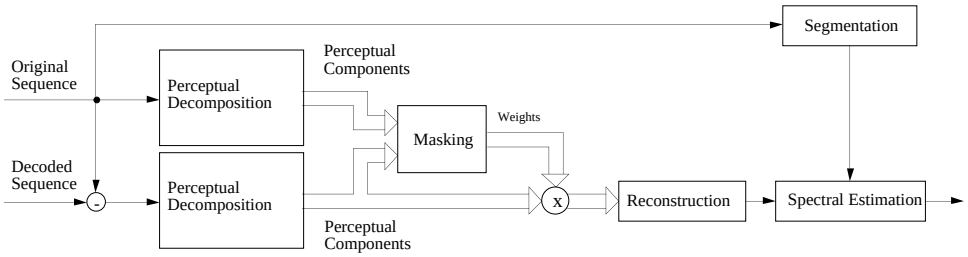

Figure 1: General block diagram of the feature quality metrics: the front end is a vision model for video that assesses pattern sensitivity, followed by spectral estimator applied on the segmented areas of the perceived distortion.

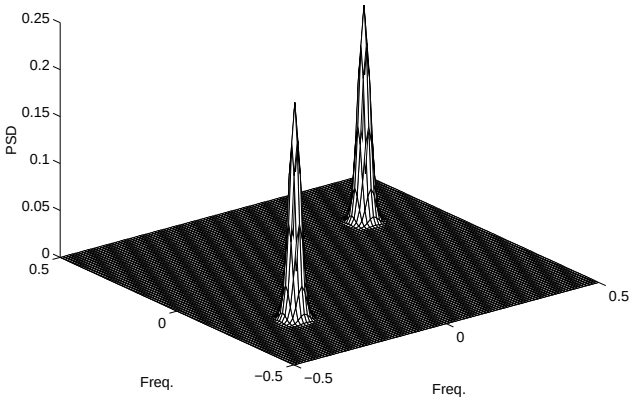

Figure 2: Estimated power spectral density of the distortion around an edge for the edge rendition sequence compressed with MPEG-2. The distortion was found to be exactly perpendicular to the actual contour.

product, in vector notation as [1]:

$$B_{n,m}(k,l) = \frac{K}{2}\left(\cos\left[\frac{\pi}{16}\begin{pmatrix}2n+1\\2m+1\end{pmatrix}\cdot\begin{pmatrix}k\\l\end{pmatrix}\right]+\cos\left[\frac{\pi}{16}\begin{pmatrix}2n+1\\2m+1\end{pmatrix}\cdot\begin{pmatrix}k\\-l\end{pmatrix}\right]\right),$$

where $\cdot$ denotes the dot product. The above formulation shows the appearance of a signal in the directions $(k,l)^T$ and $(k,-l)^T$ that are conjugate. This causes the appearance of a noise, correlated with the signal, in the conjugate direction, i.e. the impression of a contour perpendicular to the actual one.

The false contour can be modeled by a sum of decaying exponentials in frequency that we estimate by a principal-component spectral estimator (the MUSIC 2D method [10]). A side advantage of such a method is that it also permits to estimate the Gibbs phenomenon distortion if the coder were based on subband decomposition [7]. It is to be noted that the spectral estimator is applied on small regions around the contours (so that the signal of interest corresponds to the model).

An example of the metric estimation is shown in Fig. 2. The MUSIC spectral estimator has been applied on the output of the vision model around a contour and estimated the mosquito noise distortion. The sequence being evaluated was a synthetic test sequence meant to test contour rendition [8]. The sequence has been MPEG-2 encoded at a medium bitrate. It is to be noted that the measured distortion was exactly perpendicular to the actual contour, the orientation of which was unknown to the spectral estimator.

3.2. Blocking Effect

The annoyance of blocking effect is estimated in the following way. The distortion known as blocking effect is fairly easy to estimate as it roughly is a periodic horizontal and vertical signal. We are interested however in knowing how much of its energy is being perceived. For this, we consider all the non-diagonal bands at the output of the vision model (the filter bank that emulates the multi-resolution structure of the visual cortex is orientation-selective). A principal component method, known as the Tufts and Kumaresan (TK) algorithm [5], is then applied on the lines or columns of the resulting signal in order to estimate the very localized spectral lines that make the blocking effect. Let a_k's be the n poles estimated by the TK estimation. The power spectral density $P_b(e^{j\omega})$ of the distortion related to blocking effect can then be computed as:

$$P_b(e^{j\omega}) = \frac{1}{\prod_{k=1}^{n}|1-a_k e^{j\omega}|^2}\ .$$

An example of performance of the metric has been obtained with another synthetic test sequence, meant for this purpose. The sequence is the "blocking effect sequence" introduced in [8]. The sequence has been MPEG-2 encoded at 3 Mbit/sec. in TV resolution. Figure 3 presents the spectrum of the perceived distortion in the vertical and horizontal frequency bands (dashed line). The solid line is the reconstructed spectrum, having determined the parameters of the autoregressive model, by the Tufts and Kumaresan method. The position of the peaks is very well estimated. There is however a bias in the estimation of the peaks' amplitude, which is a known effect of the TK method. The estimated spectrum represents the amount of visible blocking distortion.

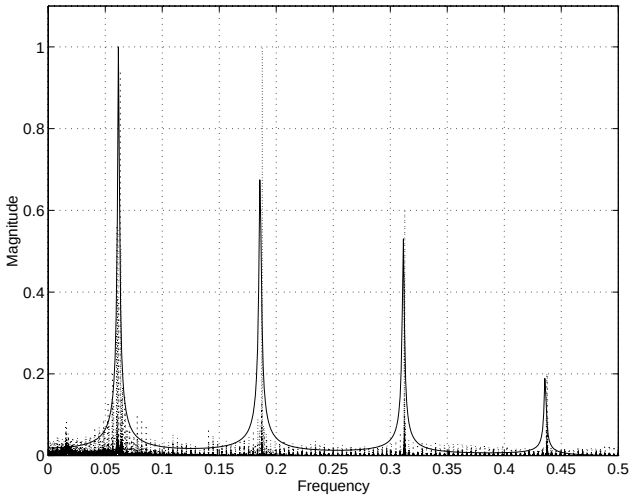

Figure 3: Example of performance of the TK estimation on a distortion signal. The normalized magnitude of the spectrum is plotted as a function of the frequency. Dashed line is the actual spectrum, solid line is the estimated spectrum.

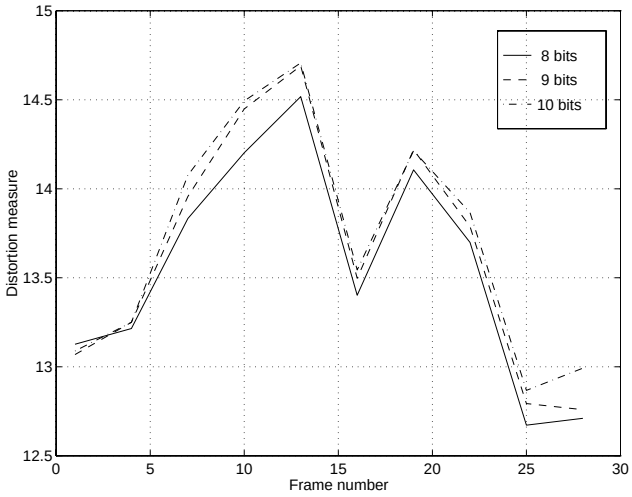

Figure 4: Temporal evolution of the texture rendition metric for the sequence Flower Garden compressed with three different precision for the DC coefficients of the DCT. Solid line is an 8 bits precision, dashed line a 9 bits precision and dot-dashed line a 10 bits precision.

3.3. Texture Distortion

The last proposed estimator addresses the visibility of textures. The method is based on the *texton* theory: a texton blob detector introduced in [4] is used. Once the detection of blob is carried out, the distortion related to them can be computed by simple pooling [7].

The blob detection and attributes computation is performed as follows: it operates on the content of a channel, i.e. the output of a filter and applies the following non-linearity:

$$\psi(x) = \tanh(\alpha x) = \frac{1 - e^{-2\alpha x}}{1 + e^{-2\alpha x}} \ ,$$

where x is a pixel of the Gabor filtered image and α is a constant set to 0.1 in our model. The non-linear function transforms the sinusoidal oscillations in the filtered images into square modulations, i.e. into blobs. A local *texture energy measure* is then computed as the average absolute deviation from the mean in small overlapping windows. Let $I(x, y)$ be a filtered image and $e(x, y)$ be the *attribute image*, i.e. the image made of texture energy measures, $e(x, y)$ is computed as:

$$e(x, y) = \frac{1}{M^2} \sum_{(a,b) in W_{xy}} |\psi(I(a, b))| \ ,$$

where W_{xy} is an $M \times M$ window centered on the pixel at position (x, y). The next stage is the segmentation itself that is performed by a pattern-clustering algorithm in the attributes' space.

As an example of this metric's performance, several compressed versions of "Flower Garden" have been obtained by varying the precision of the representation of the DCT DC coefficient, using the values of 8, 9 and 10 bits as permitted in the MP@ML mode of MPEG-2. As most information of textures is contained in the AC coefficients, one could expect few differences in texture rendition quality. This is what the MSE captures as it can be seen in Fig. 5. The texture rendition metric yields a different judgment: according to this metric, the 8-bit stream is rated as better than the two others, as it can be deduced from Fig. 4.

The prediction by the MSE and texture rendition metric are rather different. Subjective data has been collected on 5 subjects to confirm the metric's results. The subjects had to perform a 3-alternatives forced choice task: they where presented with the original sequence (always at the same place) and the three compressed sequences placed randomly, Their task was to rank order the sequences by order of visibility of the distortion on texture areas (the flowers in the sequence). The 8-bit stream has been rated as the best nine out of fifteen times, the 9-bit stream four times and the 10-bit stream only twice. The subjective data thus confirms the results of the texture rendition metric. The MSE was not able to correctly discriminate between the streams.

Looking at this result, it is important to explain why a difference in subjective quality indeed appears. Varying the DC precision should not affect the quality of textures as most of the information is in the AC coefficients. However, the coder operated in constant bitrate mode. In this case, when the DC coefficients are encoded with more bits, a smaller bandwidth is devoted to the AC coefficients that have to be quantized more coarsely, which results in a lower quality of texture rendition.

4. CONCLUSION

In this paper, an extension to the vision model presented in [9] has been introduced. The new model uses a seg-

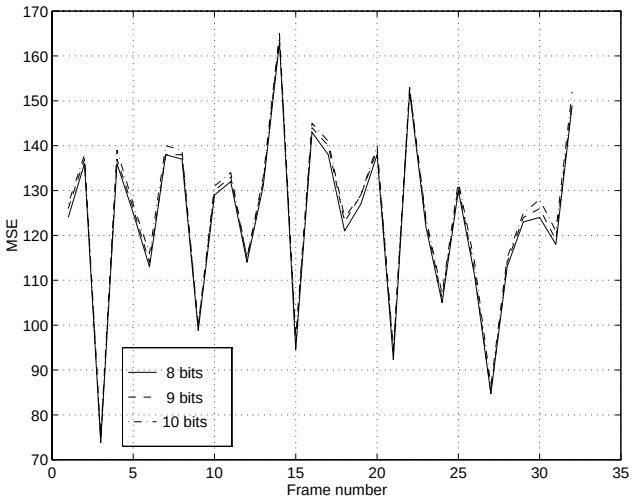

Figure 5: Temporal evolution of the MSE for the sequence Flower Garden compressed with three different precision for the DC coefficients of the DCT. Solid line is an 8 bits precision, dashed line a 9 bits precision and dot-dashed line a 10 bits precision.

mentation tool to partition the areas of the sequence into classes. A particular metric is then run on each class to estimate how the particular features of each class are rendered. Three different metrics have been introduced, one addresses contour rendition, the second estimates the annoyance of blocking effect and the third assesses the quality of texture rendition. The first two metrics are based on principal components spectral estimators, while the last one uses the texton theory.

Further results have been reported in [7] along with complete implementation details of the vision model and the metrics.

5. REFERENCES

[1] Serge Comes, Marco Mattavelli, Olivier Bruyndoncks, and Benoît Macq, "Post-Processing of Decoded Pictures by Filtering of the Unmasked Noise", *submitted to the IEEE Transactions on Image Processing*, 1995.

[2] C. P. Cressy and G. W. Beakley, "Computer-Based Testing of Digital Video Quality", *Proceedings of SPIE*, Vol. 2187, pp. 68–78, 1994.

[3] O. Egger, W. Li, and M. Kunt, "High Compression Image Coding Using an Adaptive Morphological Subband Decomposition", *Proceedings of the IEEE*, Special Issue on Advances in Image and Video Compression, Vol. 83, No. 2, pp. 272–287, February 1995.

[4] Anil K. Jain and Farshid Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters Pattern Recognition", *Pattern Recognition*, Vol. 24, No. 12, pp. 1167–1186, December 1991.

[5] R. Kumaresan and D. W. Tufts, "Estimating the Parameters of Exponentially Damped Sinusoids and Pole-Zero Modeling in Noise", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 30, No. 6, pp. 833–840, December 1982.

[6] P.H. Meehan, R.N. Hurst, M.A. Isnardi, and P.K. Shah, "MPEG Compliance Bitstream Design", *in Proceedings of the Consumer Electronics Society of the IEEE*, pp. 316–319, Chicago, IL, June 7-9 1995.

[7] Christian J. van den Branden Lambrecht, *Perceptual Models and Architectures for Video Coding Applications*, PhD thesis, Swiss Federal Institute of Technology, CH 1015 Lausanne, Switzerland, September 1996, available on `http://ltswww.epfl.ch/pub_files/vdb/`.

[8] Christian J. van den Branden Lambrecht, Vasudev Bhaskaran, Al Kovalick, and Murat Kunt, "Automatically Assessing MPEG Coding Fidelity", *IEEE Design and Test Magazine*, Vol. 12, No. 4, pp. 28–33, winter 1995.

[9] Christian J. van den Branden Lambrecht and Olivier Verscheure, "Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System", *in Proceedings of the SPIE*, Vol. 2668, pp. 450–461, San Jose, CA, January 28 - February 2 1996, available on `http://ltswww.epfl.ch/pub_files/vdb/`.

[10] H. Youdal, M. Janati-Idrissi, and M. Najim, *Modélisation paramétrique en traitement d'images*, Masson, 1994.