MOTION RENDITION QUALITY METRIC FOR MPEG CODED VIDEO

Daniele M. Costantini [†*] *Christian J. van den Branden Lambrecht* [‡**]
Giovanni L. Sicuranza [†] *Murat Kunt* [‡]

[†] Image Processing Laboratory, DEEI, Università degli Studi di Trieste, via A. Valerio, 10, 34100 Trieste, Italy
[‡] Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland

ABSTRACT

This work addresses quality assessment of motion rendition in digital video coding. Motion estimation and compensation are critical modules in video coders. A computational metric, based on a spatio-temporal model of the human visual system and of human motion sensing, is proposed and used to evaluate MPEG-2 compressed video. The metric is able to assess the quality of motion rendition and exhibits a good correlation with subjective data.
Keywords: Quality assessment, vision model, motion sensing, motion estimation, MPEG, test

1. INTRODUCTION

Video transmission systems are currently in a state of transition from a completely analog system to a digital system. The digital system which will be extensively deployed will incorporate source coders employing the MPEG compression standards. Testing of such systems is problematic as the methodology to test digital video transmission systems has not been formalized and the analog testing methodology cannot be used for digital systems. A crucial module in an MPEG encoder is the motion estimation. From an implementation viewpoint, it is the most demanding resource as the computational load and memory bandwidth involved in motion estimation are one order of magnitude higher than for any other module [2]. From a performance point of view, motion rendition is perceptually very important. This paper presents a computational metric that specifically evaluates the quality of motion rendition. It is based on a spatio-temporal multi-channel model of human vision and motion sensing. The modeling of human vision and motion sensing is addressed in Sec. 2. The metric is described in Sec. 3. Experimental results on MPEG-2 compressed video material and comparison with subjective data are presented in Sec. 4. Eventually, Sec. 5 concludes the paper.

2. MODELING MOTION SENSING

A spatio-temporal model of human vision has been developed for the framework of video coding and presented in [3, 5]. The model is based on the following properties of human vision:

- The visual system represents the information by contrast and not by absolute light level.
- The responses of the neurons in the primary visual cortex are band-limited. The human visual system has a collection of mechanisms or detectors (termed channels) that mediate perception. A channel is characterized by a localization in spatial frequency, spatial orientation and temporal frequency. The responses of the channels are simulated by a three-dimensional filter bank.
- In a first approximation, the channels can be considered to be independent. Perception can thus be predicted channel by channel without interaction.
- Human sensitivity to contrast is a function of frequency and orientation. The *contrast sensitivity function* (CSF), quantizes this phenomenon, by specifying the detection threshold for a stimulus as a function of frequency.
- Visual masking accounts for inter-stimuli interferences. The presence of a background stimulus modifies the perception of a foreground stimulus: masking corresponds to a modification of the detection threshold of the foreground according to the local contrast of the background.

The working model described in [3] incorporates the above described considerations of visual perception. The filter bank used in the model decomposes the data according to 5 spatial frequency bands (centered at 0, 2, 4, 8 and 16 cycles per degree (cpd)), 4 orientation bands (centered at 0, $\pi/4$, $\pi/2$ and $3\pi/4$) and 2 temporal frequency bands, termed the *sustained* and *transient* mechanisms. An estimate of the CSF to coding noise, based on an excitatory-inhibitory formulation, has been obtained by psychophysical experiments [3]. The model of masking used is the non-linear transducer introduced in [1].

Watson and Ahumada proposed in [6] a model of the motion sensor considering the following fundamental properties of human motion sensing:

- Humans perceive speed and direction of a movement.
- Motion is local: humans are able to discriminate between objects moving at different speeds and with different directions in a scene.
- Motion perception is dependent on spatial frequency.
- Motion detection is direction-selective at high temporal frequencies: the threshold contrast for a stimulus moving in one direction is unaffected by a similar stimulus moving in the opposite direction. At lower temporal frequencies, considerable subthreshold summation appears, showing that motion detection is non direction-selective.
- The contrast threshold necessary to detect a moving stimulus is equal to the detection threshold for the stimulus.

* Visiting student at EPFL-LTS
** Corresponding author, Email: vdb@lts.de.epfl.ch

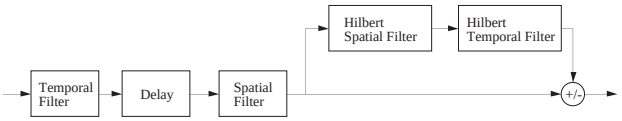

Figure 1: Block diagram of the model of a motion sensor.

Their model of the motion sensor is illustrated in Fig. 1. The first block of the sensor is a temporal filter that accounts for the global temporal contrast sensitivity of the eye. The signal is then delayed so as to ensure a causal system at the end (some of the following building blocks have a non causal response). The delayed signal is filtered by a spatial filter, the response of which is the profile of a spatial mechanism. The output of the spatial filter is then divided into two separate paths. The first path, the main path, is unprocessed, while the second undergoes a filtering operation by a spatial and a temporal Hilbert filter. The second path is said to be the quadrature path, as the signal is phase-shifted by $\pi/2$ with respect to the main path. The main and quadrature paths are then added (to compute rightward motion) or subtracted (for leftward motion) to form a direction-selective linear motion sensor.

In this work, a modification of the Watson and Ahumada's motion sensor is proposed. The sensor is adapted to the multi-resolution structure of the model described in [3]. The modifications are the following: The temporal filter no longer is the temporal contrast sensitivity function but the profile of a temporal mechanism, so as to fully match the multi-resolution structure of the model (in the spatial and temporal dimensions). For sustained mechanisms, motion sensing is non direction-selective, hence the quadrature path featuring the Hilbert transformer is needless. Transient mechanisms are direction-selective, so the Hilbert transformer is used with a temporal Hilbert filter that matches the front-end temporal filter.

3. THE METRIC

The above considerations are used to build a computational motion detector that is able to estimate motion rendition artifacts. In the context of test and evaluation of video codecs, one would be interested in predicting how good of a job a coder did in representing motion. The goal is to build a metric that takes as input an original video sequence, a compressed/decompressed version of the precedent and assesses the quality of motion rendition in the decompressed sequence. The block diagram of the proposed tool, named *motion rendition quality metric* (MRQM) is presented in Fig. 2.

The front-end of the metric is the vision model as described in [3]. The coding distortion is computed as the difference between the decompressed sequence and the original one. The original and distortion sequences are decomposed into perceptual components by the three-dimensional filter bank. The CSF and masking function are then used to compute threshold contrasts for every pixel of the perceptual components of the distortion. The data is then expressed in *just noticeable differences* by dividing the values with the detection thresholds. At this stage, processing differs for the sustained and transient channels. The transient channels are mapped onto opponent-signals to account for opponent-motion energy. The box denoted opponent-motion energy sensors thus realizes the delay operation followed by a separation into main and quadrature paths for each transient channel. The quadrature path is filtered by the spatio-temporal Hilbert filter, then added and subtracted to the main path to estimate rightward and leftward motion. The processing measures the motion energy of the rectified signal in opponent directions. The data is now to be gathered to simulate the next higher-order elaborations done by the visual cortex.

Watson and Eckert addressed this problem [7]. They showed that, once the detection of motion and the sensing of motion direction are done, further processing of the cortex can be modeled by the sensing of motion gradients. They showed that the output of motion sensors undergoes a subsequent pooling, performed over a large extent. This pooling actually takes the form of a spatial excitatory-inhibitory pooling over a Gaussian-shaped area. It is performed on both the sustained and transient channels and constitutes the last processing stage of MRQM. The data of Watson and Eckert [7] are used to implement the pooling operation. The metric finally outputs a distortion measure for each channel.

4. EXPERIMENTAL RESULTS

As an example of experiment, the study of motion rendition quality as a function of the search window dimension is presented. More extensive results are presented in [4]. In this experiment, the sequence Basket Ball has been compressed at a rate of 6 Mbit/sec., as interlaced video material, using a constant group of picture structure of 12 frames with 2 B-pictures between each P-picture. The video buffer verifier size was set to its maximum allowed size. The coder operated in constant bitrate (CBR) mode. Various streams have been obtained by varying the dimension of the search window. The same search area has been selected for every frame, i.e. P or B frames have the same search area. Psychophysical tests have then been carried out to confirm the results obtained with MRQM.

Two types of results are now presented. In one case, the metric is run on the decoded streams. The streams thus incorporate all effects of encoding, including motion compensation and quantization of the displaced frame difference (DFD). In the other case, the metric is computed on the prediction frames, i.e. the frames obtained at the output of the motion estimation algorithm, *before* motion compensation. This permits to study motion estimation artifacts only, without any effects of motion compensation or DFD quantization. Such results are presented in Fig. 3, top (measurements on the decoded frames) and in Fig. 3, bottom (measurements on the prediction frames) for four search window dimensions, namely 9×9, 15×15, 31×31 and 63×63. In such graphs, the distortion measure is plot channel by channel. Channels are ordered first by spatial frequency, then by orientation, i.e. channels 1 to 4 correspond to a spatial frequency of 2 cpd and orientations of

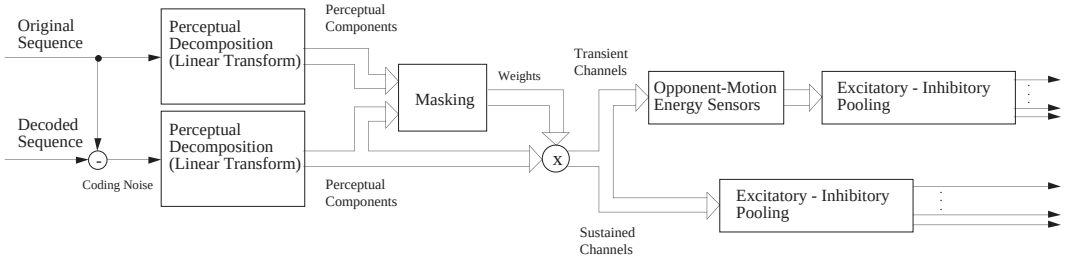

Figure 2: Block diagram of the motion rendition quality metric.

0, $\pi/4$, $\pi/2$ and $3\pi/4$. The next four channels correspond to a spatial frequency of 4 cpd, etc. Only transient channels are represented here as they account for most of the distortion in motion rendition [4].

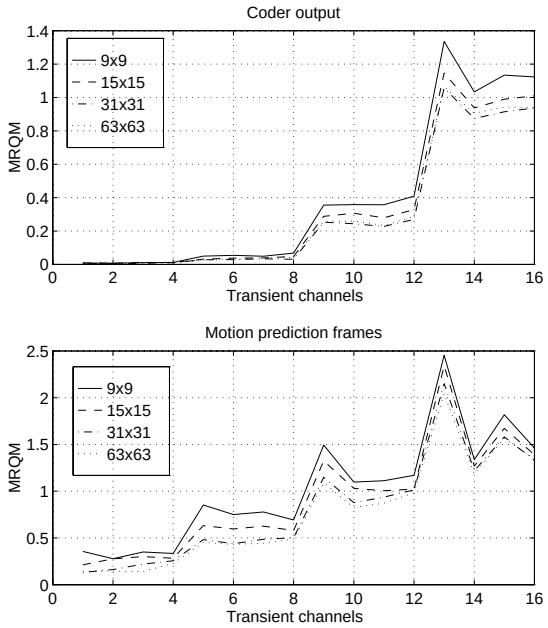

Figure 3: Top: MRQM measurements on decompressed frames of Basket Ball compressed at 6 Mbit/sec. with different motion estimation search windows. Bottom: MRQM measurements on the prediction frames of Basket Ball compressed at 6 Mbit/sec. with different motion estimation search windows.

The graphs are interesting in many aspects. If one looks at the MRQM output on the motion prediction frames (Fig. 3, bottom), it can be seen than the distortion in motion decreases as the dimension of the search window increases. This is normal, as the prediction made by full search can only be better as the window size increases. Patterns of high spatial frequency are more difficult to render (MRQM increases with spatial frequency). This is also due to the fact that the motion prediction frames are subject to the quantization of the reference frames. Therefore, distortion increases with spatial frequency as quantization of intra macroblock increases as well.

When looking at the results computed on the decoded frames, one can notice that the ordering of the streams, rated by MRQM, changes. The stream compressed with the largest search window is not the best one. This might seem awkward but is justified by the variable length coding (VLC) of MPEG: motion vectors are encoded in the bitstream using VLC tables. Several tables can be used depending on the vectors' maximal length. Hence, whatever the actual length of the motion vector is, it is encoded using a table dependent on its maximal allowed size. Therefore, streams compressed with large search windows use more bandwidth to represent motion vectors. The consequence of this is that a smaller bandwidth is devoted to the DFD that is quantized more coarsely. The use of a large search window may result in a sequence that has a lower visual quality. Such an effect is indicated by MRQM but is not captured by the peak signal to noise ratio (PSNR) as it will be shown below and as it has been observed in [2].

It is thus interesting to study which search window is best for a given sequence and see if MRQM is able to predict the responses of human observers. Such a result is presented for Basket Ball in Fig. 4, top. The graph presents the quality rating by MRQM and PSNR for Basket Ball as a function of the search window. In this graph, the output of MRQM has been pooled over channels with a Minkowski summation of exponent 4 and expressed on a logarithmic scale. The PSNR curve shows a slight knee in quality around a window dimension of 50×50. The MRQM curve, on the contrary, exhibits a maximum at a window dimension of 31×31. This result has been validated by subjective data collected on 4 human observers with a three-alternatives forced choice task: the subjects were presented 4 sequences simultaneously. The original sequence was always presented at a known place. The three other sequences were compressed sequences. The subjects were asked to choose the sequence that had the lowest distortion. The presentation time was unlimited. Several trials were performed for each subject (between 30 and 50). The data has been averaged over subjects to deduce a rank ordering. The validation is done as follows: the MRQM and PSNR outputs are plot as a function of the subjective rank ordering in Fig. 4, bottom for the four sequences considered

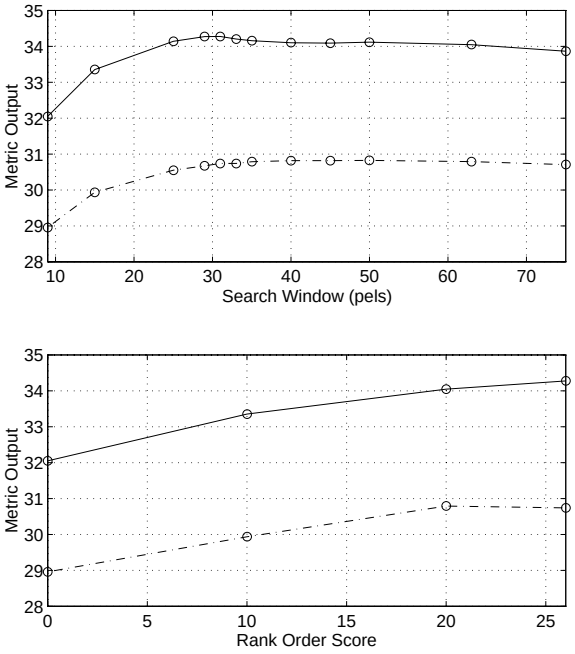

Figure 4: Top: study of motion rendition quality assessment by MRQM (solid line) and PSNR (dashed line) for Basket Ball as a function of the search window. Bottom: metric output versus subjective rank ordering for compressed versions of Basket Ball, varying the search window dimension. MRQM is the solid line, PSNR the dashed line

in Fig. 3, namely search windows of 9×9, 15×15, 31×31 and 63×63. Moreover, the horizontal distance between points of the graph is dependent on the discrimination capability. This permits to account for noise in the subjective data. The abscissas in the graph thus correspond to the ordering 9×9, 15×15, 63×63, 31×31, i.e. the rank ordering in quality given by the subjects, with a horizontal spacing that accounts for the performance in discriminability between the sequences. The graph thus plots the metric as a function of the subjective data to see how much correlated they are.

For a metric to pass this test, it needs a monotonic relationship between its output and the rank ordering. The bottom of Fig. 4 presents the curve for PSNR and MRQM. It can be seen that PSNR is not able to predict the subjective data in some cases (namely for large search window) and an inversion in ordering appears for the last two sequences. PSNR indeed predicts that the sequence compressed with a search window of 63×63 looks better than the one using a window of 31×31. MRQM, on the contrary, is able to predict the subjective data. It exhibits a behavior that is nearly linear with the subjective rank ordering, which shows a very good correlation with this data. It is to be noted that MRQM predicts that the best search window for Basket Ball is 31×31, which is confirmed by human observers.

5. CONCLUSION

This paper presented a computational metric that is devoted to assessing the quality of motion rendition in digital video coding. The metric, termed MRQM, is built on top of a spatio-temporal multi-channel vision model. It features a modeling of the motion sensors as they are thought to mediate motion sensing and discrimination of moving objects. The model of the motion sensor differs for transient and sustained channels. The sustained motion sensor is non direction-selective. The transient motion sensor, on the contrary, is direction-selective. Both detectors undergo an excitatory-inhibitory pooling that models further elaborations of the visual cortex. The metric has been used to study the relationship between quality of motion rendition and dimension of the search window in MPEG-2 coding. MRQM predicts quality as it should, discriminating or ordering compressed sequences as human observers would do it. The metric prediction were compared with subjective data and exhibited a good correlation with the data, which is not the case for the PSNR.

6. REFERENCES

[1] Gordon E. Legge and John M. Foley, "Contrast Masking in Human Vision", *Journal of the Optical Society of America*, Vol. 70, No. 12, pp. 1458–1471, December 1980.

[2] Marco Mattavelli, Christian J. van den Branden Lambrecht, Didier Nicoulaz, and Carlos López Fernández, "Low Complexity Motion Estimation Heuristic for MPEG-2 Systems", *IEEE Transactions on Circuits and Systems for Video Technology*, 1996, submitted paper.

[3] Christian J. van den Branden Lambrecht, "A Working Spatio-Temporal Model of the Human Visual System for Image Restoration and Quality Assessment Applications", *in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2293–2296, Atlanta, GA, May 7-10 1996, available on `http://ltswww.epfl.ch/~vdb`.

[4] Christian J. van den Branden Lambrecht, Daniele M. Costantini, Giovanni L. Sicuranza, and Murat Kunt, "Quality Assessment of Motion Rendition in Video Coding", 1996, in preparation.

[5] 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/~vdb`.

[6] Andrew B. Watson and Albert J. Ahumada, "Model of Human Visual-Motion Sensing", *Journal of the Optical Society of America*, Vol. 2, No. 2, pp. 322–341, February 1985.

[7] Andrew B. Watson and Michael P. Eckert, "Motion-Contrast Sensitivity: Visibility of Motion Gradients of Various Spatial Frequencies", *Journal of the Optical Society of America*, Vol. 11, No. 2, pp. 496–505, February 1994.