EFFICIENT SPATIO-TEMPORAL DECOMPOSITION FOR PERCEPTUAL PROCESSING OF VIDEO SEQUENCES

Pär Lindh [†] *and Christian J. van den Branden Lambrecht*

Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
vdb@lts.de.epfl.ch
http://ltswww.epfl.ch/~vdb/

ABSTRACT

This paper presents a vision model for moving pictures. The model is an extension of a normalization model by Teo and Heeger. It accounts for normalization of the cortical receptive field responses and inter-channel masking. The model is compared with a simpler vision model for video by presenting results on quality assessment of MPEG compressed video.
Keywords: Quality assessment, vision model, MPEG, test

1. INTRODUCTION

There has been a strong interest over the past years in vision models and applications of vision science in image processing. This interest comes from the fact that the modeling of vision permits to predict the visibility of patterns, which is of interest to many applications such as image quality assessment [8], image coding [13] or display device design. Vision models have been mainly developed for still picture applications. Recently, a model for moving pictures has been proposed [10] for applications in a video coding framework [12]. The model relies on a simple modeling of the cortical receptive fields. Vision scientists have proposed nonlinear models of early vision that accounts for contrast normalization in the cortical neurons responses [4, 8, 9]. They showed that such models perform better compared to simpler modeling of early vision. This work presents an extension of such a still-picture model so as to be applicable to moving pictures. The model by Teo and Heeger [8], that is being extended here, is reviewed in Sec. 2. Section 3 presents the spatio-temporal normalization model and Sec. 4 presents results of the distortion metric obtained with the model. Finally, Sec. 5 concludes the paper.

2. THE TEO-HEEGER MODEL

Models of early vision attempt to predict the responses of the neurons in the primary visual cortex (termed area V1). Such modeling is based on the following properties:

- The visual system uses the relative contrast as the representation of the information.
- The visual information is represented at various scales and orientations. Psychophysical experiments give evidence for the existence of separate detection mechanisms in the brain, which is confirmed by the measurements of band-pass responses from the cortical cells. The portion of the frequency domain that a mechanism covers is called a *channel*.
- The sensitivity to contrast is a function of the frequency and the orientation. This function is termed the *contrast sensitivity function* (CSF).
- The adaptation to the local contrast of the background, i.e. visual masking.

A model of visual masking that is commonly used has been proposed by Legge and Foley [5]. It sums excitation linearly over a receptive field. The formulation assumes that visual masking can only occur between two stimuli that belong to the same channel. Such modeling has been contradicted by experiments carried out by Foley and Boynton on simultaneous masking of Gabor patterns by sinewave gratings [3]. In these experiments, target contrast thresholds were measured as a function of the masker contrast, orientation, spatial phase and temporal frequency.

A very important result of their study is the effect of masker orientation on the threshold curves. A conclusion is that there exists some inter-channel masking and that this phenomenon can be significant. An illustration of the types of measurements that they obtained is given in Fig. 1. Consider two stimuli, a masker and a target. Let C_{T0} denote the detection threshold of the target measured in the absence of a masker (i.e. given by the CSF), C_T be the detection threshold of the signal in the presence of a masker and C_M be the contrast of the masker. If the masker and the target have the same orientation, the diagram of masking looks like the one depicted in Fig. 1, left hand side. There is a region, for $C_M < C_{T0}$, where the masker has no influence on the perception of the target. For $C_M > C_{T0}$, the detection contrast for the target grows exponentially with the contrast of the masker (i.e. detection of the target is harder). When the contrast of the masker and the target are very close one to the other, the curve presents a dipper. This means that, at this value of contrast, the masker actually helps detection of the target (this is known as the facilitation effect). When the target and the masker differ

[†] Visiting student from Linköping Institute of Technology

in orientation, the detection curve looks as the one depicted in Fig. 1, right hand side. The difference with the previous case is that the curve does not present a dipper around $C_M = C_{T0}$, i.e. there is no facilitation effect.

On the basis of his modeling of the cat cortical cell's response [4] and the data from Foley and Boynton, Heeger proposed with Teo a fidelity metric for still pictures [8, 9]. The model is based on four building blocks: a front end linear transform, a squaring of the transform coefficients, a normalization stage and a detection stage.

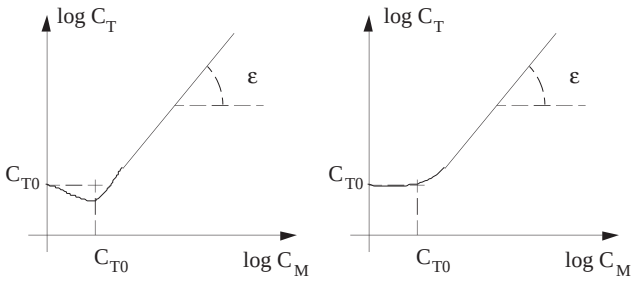

Figure 1: Detection contrast curve for a target in the presence of a masker. Left hand side: the masker and the target have approximately the same orientation. Right hand side: the masker and the target have different orientations.

The linear transform stage decomposes the image into perceptual channels. It is implemented with hexagonally sampled quadrature mirror filters or cosine filters in [9] or the steerable pyramid [7] from Simoncelli *et al.* in [8]. The coefficients at the output of the linear transform are squared and normalized. Since the output coefficients of the transform linearly increase with the input magnitude, it is desirable to restrict the output to a certain dynamic range (as it is the case in an actual system like the cortex). Let A^θ be a coefficient of the output of the linear transform having orientation θ. The normalized output for the coefficient, R^θ, is computed by Eq. (1):

$$R^\theta = k \frac{\left(A^\theta\right)^2}{\sum_\phi \left(A^\phi\right)^2 + \sigma^2} \ , \qquad (1)$$

where ϕ ranges over all orientations, k's is a global scaling constant and σ's a saturation constant. Pooling is only performed across orientations and not across spatial frequencies as inter-channel masking is restrained to channels having the same frequency [3]. The values for constants k and σ have been obtained by fitting the model to Foley and Boynton's data. In this formulation, the presence of the dipper in the masking curve can be quite simply explained. If the masker and the target have approximately the same orientation, a contribution at that orientation appears both in A^θ and one A^ϕ, decreasing the value of R^θ. On the contrary, when both signals have different orientations, there is no simultaneous contribution to A^θ and an A^ϕ.

Finally, a detection mechanism predicts how different two images may look. Assume that R_o is a vector gathering all the sensors outputs computed by Eq. (1) for an original image and R_d is the equivalent vector for a distorted version of the considered image. The distortion measure is computed as the squared error norm of the difference between R_o and R_d as: $\Delta R = |R_o - R_d|^2$.

3. SPATIO-TEMPORAL NORMALIZATION MODEL

The Teo-Heeger model has been extended to account for spatio-temporal perception. In order to do this, a spatio-temporal decomposition has to be obtained along with values for the normalization equations constants. The decomposition is chosen to be time-space separable as in [10, 11]. This can be done if the contrast sensitivity function is kept non separable in space and time to account for spatio-temporal interactions. The spatial filterbank is kept as in the Teo-Heeger model, i.e. it is the steerable pyramid [7]. The pyramid splits the data into four spatial frequency bands and four orientation bands.

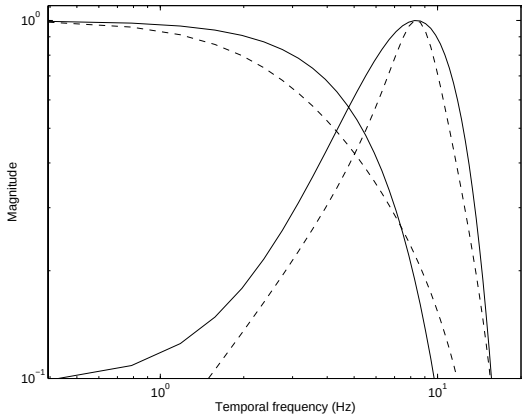

Figure 2: Comparison of the magnitude of the frequency responses of the temporal bank of [10] (solid line) and the proposed IIR filterbank (dashed line).

A temporal filter bank is proposed. The main constraint that has been imposed when designing it is low delay. Some applications that would use a perceptual decomposition may require very fast response from the model. Consider for example a buffer regulation scheme for an encoder. The quantization parameter has to be adapted as a function of the desired quality and occupancy of the output buffer. Such decisions have to be made as quickly as possible. For this reason, no subsampling operation is performed along the temporal direction. This was chosen so as to have as much degrees of freedom as possible in designing the temporal filterbank. The bank also has to approximate the two mechanisms of temporal vision, that are termed *sustained* and *transient*. Doing so with finite impulse response filters (FIR filters) may require filters having as much as 6 to 8 taps, which is a too long delay. Therefore infinite impulse response (IIR) filters have been chosen to implement

the temporal decomposition. A maximum delay of 3 frames has been imposed. Such a delay is reasonable as many implementations of actual coders, namely MPEG coders, have such delays.

The filters have been designed by minimizing the difference between their frequency response and the response of the filter used in [10] in a least square sense. The optimization procedure yielded a low pass filter that has one pole and one zero. This filter approximates the sustained mechanisms. The transient mechanism is approximated by a filter with 3 poles and 2 zeroes. Finally, a third filter is designed to yield the high pass residue (so as to have an invertible transform). This filter added to the sum of the two others yields a flat response. All filters are listed in [6]. The low pass and band pass filters are presented in Fig. 2. The values for the constants k's and σ's have been obtained by simulating a target at threshold contrast. For this stimuli, the sum of the output of all sensors should equal unity. The contrast sensitivity function experimentally obtained in [11] has been used for this purpose.

4. PERCEPTUAL DISTORTION METRIC

The proposed model has been used to design a metric for moving pictures termed the *normalization fidelity metric* (NVFM). In this metric, the output of the detection stage ΔR is further mapped onto the 1 to 5 quality scale defined by CCIR Rec. 500 [2]. In this scale, 1 is the worst quality and 5 the best. The mapping uses the following function, relating the error measure to the quality index Q:

$$Q = \frac{5}{1 + N\Delta R} ,$$

where N ensures a mapping between 1 and 5 and is computed on the basis of the vision model as in [12].

Results of quality assessment on compressed video for broadcasting are now presented. The considered coder is the MPEG-2 standard operating in MP@ML (main profile, main level) and HP@ML (high profile, main level). Two classical test sequences for broadcasting applications have been used for the simulations. They are *Mobile & Calendar* and *Basket Ball*. The sequences have been encoded with the software simulator of the test model 5 of MPEG-2 supplied by the MPEG Software Simulation Group as interlaced video, with a constant group of pictures structure of 12 frames and 2 B-pictures between every P-picture. The video buffer verifier size was set to its maximum allowed size. The dimension of the search windows for motion estimation was 15 pixels for P-frames, 7 pixels for backward motion estimation in B-frames and 3 pixels for forward motion estimation in B-frames. The coder operates in constant bitrate (CBR) mode. Coding has been performed on the range of bit rates that MPEG-2 typically addresses. The coding quality of the sequences Mobile & Calendar and Basket Ball has been assessed by NVFM and is compared with the moving pictures quality metric (MPQM) introduced in [12]. The results are respectively presented in Fig. 3 and Fig. 4. The general behavior of the curves indicates a rapid increase in quality at low to medium bitrates and a saturation at higher bitrates. The shape of the NVFM curves is significantly different from the MPQM curves in that their dynamic range spans a much larger range in quality. The first portion of the NVFM curves exhibits a steep limb. According to NVFM, the increase in quality in the lower range of bitrate is very fast, and a slight increase of bandwidth can result in a very significant increase in quality. It is interesting to note that saturation occurs roughly at the same bitrate value for both metrics.

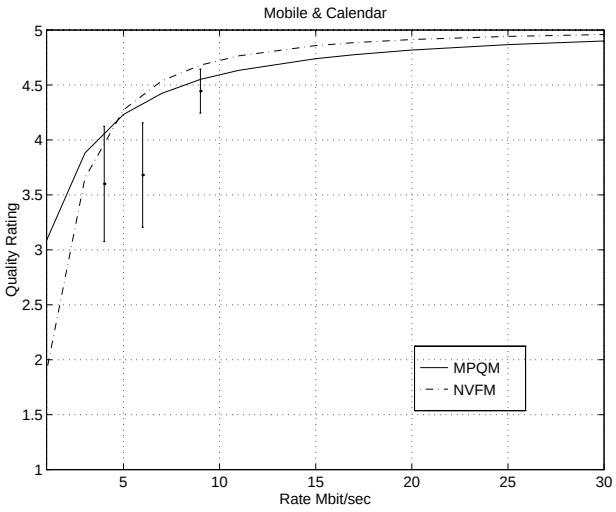

Figure 3: Comparison of NVFM, MPQM and subjective data for Mobile & Calendar as a function of the bitrate.

The metric is also compared with some available subjective data. The data has been collected by the research center of RAI, Italy [1] and consists of subjective rating of compressed video by human observers. The data has been collected according to CCIR Rec. 500-3. The method is a double stimulus continuous quality scale (DSCQS). The subjects are presented pairs of sequences. Each pair consists of versions of the same sequence (i.e. the sequences are chosen between the various compression ratios and the original). The observer has to assess the quality of *both* sequences on a scale that is similar to the CCIR quality scale. The subjective data has thus to be adapted to the purpose of this experiment. Both the original and the compressed sequences are given a vote in the DSCQS task. The perceptual metrics, on the contrary, try to predict how different two sequences may look. The output of the metric is always a distance between the distorted sequence and the original. Hence the data from [1] has been used as follows. Each result has been normalized with respect to the original and the distance between the two subjective votes used to deduce an error bar.

Figure 3 presents the curves of the three metrics for the sequence Mobile & Calendar along with the tentative mapping on the subjective data. It can be seen that the subjective data is pretty noisy since the ratings at 4 and 6 Mbits/sec. are nearly identical. The metric curves show however a behavior that is consistent with the data. Figure 4 presents the same results for the Basket Ball sequence along with the performance of the ITS metric. This

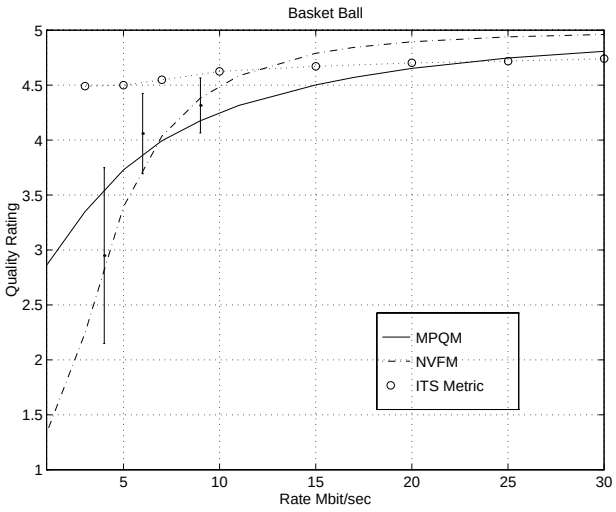

Figure 4: Comparison of NVFM, MPQM and subjective data for Basket Ball as a function of the bitrate.

metric is the only alternative metric for moving pictures and is described in [14]. The subjective data seems less noisy in this case. This data show particularly well the two phenomena discussed before, namely the increase of perceived quality with the bandwidth in the lower range of bitrates and a saturation effect at higher bitrates. Both metrics exhibit a behavior that is consistent with the data. The NVFM realizes a better match with the data than MPQM. In particular it accounts better for the rapid increase of quality below 7-9 Mbits/sec. As the NVFM models more aspects of vision than MPQM (namely inter-channel masking and normalization), one could expect that its performance would be better. The ITS metric on the contrary is not consistent at all with such data as shown in Fig. 4.

5. CONCLUSION

A first vision model for moving pictures applications has been proposed recently. In this paper, a more advanced model is presented. The new model offers several advantages compared to the previous one, namely a better modeling of the cortical cells responses and a more efficient implementation. The proposed model is an extension of the work by Teo and Heeger. Two quality metrics for moving pictures, derived from the two different models are used to assess the coding quality of MPEG-2 video streams. The new model proved to yield a better quality rating than the old one.

6. REFERENCES

[1] M. Ardito, M. Barbero, M. Stroppiana, and M. Visca, "Compression and Quality", in L. Chiariglione, editor, *Proceedings of the International Workshop on HDTV 94*, pp. B–8–2, Torino, Italia, October, 26-28 1994. Springer-Verlag.

[2] CCIR, "Method for the Subjective Assessment of the Quality of Television Pictures", *13th Plenary Assembly, Recommendation 500*, Vol. 11, pp. 65–68, 1974.

[3] John M. Foley and Geoffrey M. Boynton, "A New Model of Human Luminance Pattern Vision Mechanisms: Analysis of the Effects of Pattern Orientation, Spatial Phase and Temporal Frequency", in T. A. Lawton, editor, *SPIE Proceedings*, Vol. 2054, Computational Vision Based on Neurobiology, pp. 32–42, 1994.

[4] David J. Heeger, "Normalization of Cell Responses in Cat Visual Cortex Visual Neuroscience", *Visual Neuroscience*, Vol. 9, pp. 181–197, 1992.

[5] 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.

[6] Pär Lindh, "Perceptual image sequence quality metric using pixel domain filtering", Master's thesis, Linköping Institute of Technology, Linköping, Sweden, April 1996.

[7] Eero P. Simoncelli, William T. Freeman, Edward H. Adelson, and David J. Heeger, "Shiftable Multiscale Transforms", *IEEE Transactions on Information Theory*, Vol. 38, No. 2, pp. 587–607, March 1992.

[8] Patrick C. Teo and David J. Heeger, "Perceptual Image Distortion", *in Proceedings of the International Conference on Image Processing*, pp. 982–986, Austin, TX, November13-16 1994.

[9] Patrick C. Teo and David J. Heeger, "Perceptual Image Distortion", *in Proceedings of SPIE*, Vol. 2179, pp. 127–141, San Jose, CA, February 1994.

[10] 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`.

[11] Christian J. van den Branden Lambrecht and Murat Kunt, "Characterization of Human Visual Sensitivity for Video Imaging Applications", *Signal Processing*, 1996, submitted paper.

[12] 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`.

[13] Andrew B. Watson, "Efficiency of an Image Code Based on Human Vision", *Journal of the Optical Society of America*, Vol. 4, No. 12, pp. 2401–2417, December 1987.

[14] Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson, Stephen D. Voran, and Stephen Wolf, "An objective video quality assessment system based on human perception", *in SPIE Human Vision, Visual Processing, and Digital Display IV*, Vol. 1913, pp. 15–26, San Jose, CA, Feb. 1993.