# A WORKING SPATIO-TEMPORAL MODEL OF THE HUMAN VISUAL SYSTEM FOR IMAGE RESTORATION AND QUALITY ASSESSMENT APPLICATIONS

*Christian J. van den Branden Lambrecht*

Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne
Switzerland
vdb@lts.de.epfl.ch

## ABSTRACT

This paper describes a spatio-temporal model of the human visual system (HVS) for video imaging applications, predicting the response of the neurons of the primary visual cortex. The model simulates the behavior of the HVS with a three-dimensional filter bank which decomposes the data into perceptual channels, each one being tuned to a specific spatial frequency, orientation and temporal frequency. It further accounts for contrast sensitivity, inter-stimuli masking and spatio-temporal interaction. The free parameters of the model have been estimated by psychophysics. The model can then be used as the basis for many applications. As an example, a quality metric for coded video sequences is presented.

## 1. INTRODUCTION

Impressive progresses have been made during the last decade in the field of still image and video processing. The importance that visual information and communication took in today's society gave birth to an increasing demand for fast and efficient ways of transmitting visual data. This evolution started with analog television and is now characterized by digital techniques for representing and conveying visual information. Video coding has been one of the most prolific fields of application and a new generation of video communication consumer products is about to be released. More recently, several applications such as image quality assessment, image enhancement or even image coding showed the need to incorporate knowledge of the only element that has not been considered much: the end user. It is now understood that imaging technology in general can highly benefit from insights of vision science.

Few connections exist between both fields tough. The objective of this work is to study the benefit that video imaging application can gain from vision science. It presents a model that incorporates the characteristics of the HVS that are relevant to image processing applications. Its major innovation is to be able to deal with *video-sequences*, i.e. it incorporates a modeling of both the spatial and temporal aspects of human vision as well as their interaction. The resulting model features a three dimensional filter bank that simulates the various mechanisms of human vision, and predicts the response of the primary visual cortex by simulating contrast sensitivity and masking. The paper is structured as follows: Section 2. presents the human visual system, Sec. 3. shows how the model is built based on such knowledge. A general architecture for the use of the model is presented in Sec. 4. Parameterization of the model is briefly described in Sec. 5. and an example of application is discussed in Sec. 6. Eventually, Sec. 7. concludes the paper.

## 2. THE HUMAN VISUAL SYSTEM

Several levels of description can be adopted to study human vision. The approach that vision science uses is the one of cognitive psychology or psychophysics. The human visual system is modeled as a system characterized by a response relating the output to input stimuli. Models are then validated by psychophysical experiments in which human subjects are asked to assess visibility of stimuli. Such modeling can be efficiently performed by considering three major aspects of vision: a multi-channel structure, contrast sensitivity and masking.

### 2.1. Multi-Channel Structure

Electro-physiological experiments performed on cells of the primary visual cortex (area V1) have shown that the response of such neurons is tuned to a band limited portion of the frequency domain [1]. Such data have been confirmed by psychophysical experiments [2], giving evidence that the brain decomposes the spectra into so-called *perceptual channels* that are bands in spatial frequency, orientation and temporal frequency. Each channel can thus be seen as the output of a filter, which is characterized by a response tuned to a specific spatial frequency, orientation and temporal frequency. The profile of the channels is very close to Gabor functions [3], which may be thought of an efficient representation of visual information since they are the most compact functions both in space/time and frequency.

The number of channels that are involved in human vision has been studied by psychophysics. Its is generally admitted that temporal vision is governed by two mechanisms, termed *transient* and *sustained* [4, 5]. The first mechanism is sensitive to moving patterns, whereas the second is responsible for perception of still or slowly moving images. As far as spatial vision is concerned, it seems that there are about five bands in spatial frequency, performing an octave-band division of the frequency axis, and roughly four to eight equal-width orientation bands. Four orientation bands are used in this model.

### 2.2. Contrast Sensitivity

The response of the human eye varies as a function of frequency. This is commonly referred to as *contrast sensitivity*. More precisely, a signal is only detected by the eye if its contrast is greater than a certain threshold defined as the *detection threshold*. The detection threshold varies as a function of frequency. The sensitivity is defined as the inverse of the detection threshold and is thus a function of frequency as well. The term *contrast sensitivity function* (CSF) is usually used to denote this function. It indicates the

 

contrast that a stimulus at a specific spatio-temporal frequency should have to have a probability of being detected of 0.5. In other words, the CSF is the curve of the inverse of the detection threshold as a function of frequency.

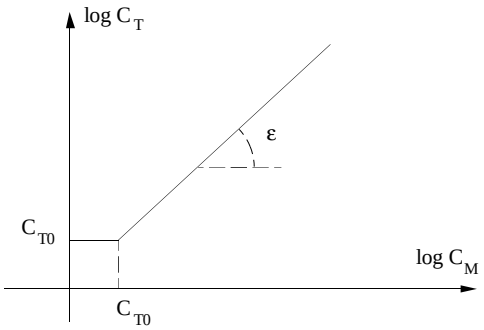


**Figure 1. Model of masking.**

### 2.3. Masking

The CSF can account for the perception of a single stimulus. However, interactions appear when several stimuli are present. In a first approximation, it is commonly considered that interference between two stimuli can only occur if they are contained in the same channel. Such interference results in a modification of the detection threshold of one stimulus due to the presence of the other. A common model of this phenomenon, termed *masking*, is a non-linear transducer as illustrated in Fig. 1. Consider two stimuli, a target and a masker. Let $C_{T0}$ be the contrast detection threshold for the target as given by the CSF, $C_M$ be the contrast of the masker and $C_T$ be the actual detection threshold for the target in the presence of the masker. The latter is determined as:

$$C_T = \begin{cases} C_{T0} & \text{if } C_M < C_{T0} \\ C_{T0} \left(\frac{C_M}{C_{T0}}\right)^{\varepsilon} & \text{otherwise.} \end{cases}$$

Therefore, when $C_M$ is greater than $C_{T0}$, the actual detection threshold, $C_T$, increases as a power of $C_M$. This dependency, plotted in a log-log graph, is a straight line of slope $\varepsilon$.

It is now known that there are some masking effect across channels, although this is a secondary phenomenon. Some recent spatial vision models [6, 7] now account for this effect as well. For the time being, since the goal is to validate the spatio-temporal model and to limit computational complexity, this phenomenon is neglected in the model.

## 3. BUILDING THE MODEL

The behavior of the human visual system can thus be modeled by cascading a three-dimensional filter bank and the non-linear transducer that models masking. The filter bank used in this model is separable in spatial and temporal frequency directions. It features 17 spatial filters and 2 temporal filters. The low-low spatial filter is isotropic. The others are tuned in four orientations (0, $\pi/4$, $\pi/2$ and $3\pi/4$ radians), and four frequency bands (centered at 2, 4, 8 and 16 cycles per degree (cpd)). The spatial filter bank is illustrated in Fig. 2 and the temporal bank in Fig. 3.

A further aspect that has to be modeled is the spatio-temporal interaction of human vision. It is well known that spatial and temporal perception are not separable [4]. Some

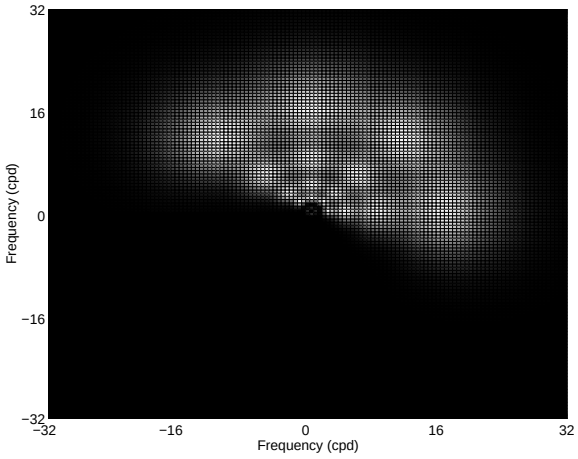


**Figure 2. The spatial filter bank, featuring 17 filters (5 spatial frequencies and 4 orientations). The magnitude of the frequency response of the filters are plotted on the frequency plane. The lowest frequency filter is isotropic.**

authors attribute this dependence to a variation in the filter positions within the spatio-temporal frequency domain, whereas others explain the phenomenon by a variation of the filter gains. Recent studies [5] gave more evidence for the latter hypothesis, permitting the use of a filter bank separable along the spatial and temporal frequency direction. The variation in the filter gain is then modeled at the level of the CSF, that will be non-separable and account for the spatio-temporal interaction. Burbek & Kelly [8] proposed an interesting modeling of the non separable CSF based on an excitatory-inhibitory formulation. This approach has been chosen to model the spatio-temporal interaction. The formulation expresses the CSF as the difference between two separable mechanisms, denoted *excitation* and *inhibition*, which permits to parameterize the whole CSF with a limited number of free parameters.

The above described Gabor filter bank may not be sufficient for some applications. Such a decomposition is not complete and cannot span the whole frequency domain. This may cause two problems: first of all, it will not be possible to reconstruct the data, which may be of interest in some applications. Secondly, some area of the spectrum will be attenuated too much and the representation of the data will suffer from scalloping. To overcome these limitations, a second filter bank is introduced. This bank is an approximation of the Gabor filter bank but offers perfect reconstruction and is an extension of the one proposed in [9]. The advantage of such a bank is that it allows reconstruction of the data after processing within the subbands and does not have any scalloping effect.

## 4. PERCEPTUAL PROCESSING OF VIDEO SEQUENCES

The structure of the working model is illustrated in Fig. 4. A video coding framework is assumed, into which an original sequence is considered along with a distorted version of it. The error sequence is computed by subtracting the original sequence from the decoded one. Then the original and the error sequences are decomposed by the filter bank. If reconstruction is needed, the perfect reconstruction (PR)

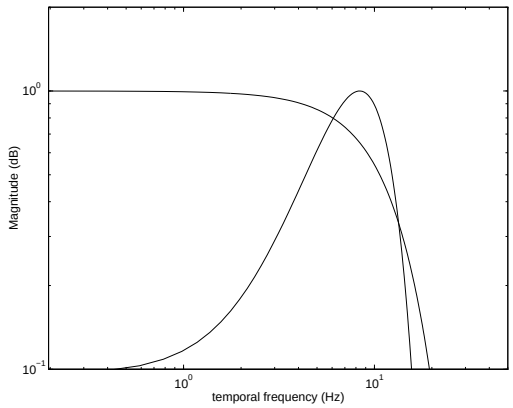


**Figure 3. The temporal filter bank accounting for two mechanisms: one low pass (the sustained mechanism) and one band pass (the transient mechanism). The frequency response of the filters is plotted as a function of temporal frequency.**

bank is used, otherwise the Gabor bank is chosen. As the original sequence will act as a masker with respect to the distortion, the non linear transducer is used to compute, pixel by pixel, channel by channel the detection threshold of the error. The error signal is then multiplied by the inverse of the detection threshold to express data in *just noticeable differences* (*jnd*'s) or *units above threshold*. This processing of the data predicts the response of the neurons in area V1.

A final stage has then to be added according to the desired application and to account for higher levels of cognition. Several approaches are possible: once the prediction of the response of the primary cortex is assessed, the data can be reconstructed and processed by classical image sequence algorithms (e.g. spectral estimation). This approach is the less likely, however. A second possibility consists in directly processing the perceptual components. In that case, the data will be pooled over the channels according to some summation rule designed for the targeted application and accounting for later stage of processing by the brain. Another approach consists in processing the perceptual components and then reconstructing the data with the PR filter bank.

## 5. PARAMETERIZATION OF THE MODEL

The model has been parameterized by means of psychophysics. The goal of the experiments was to measure the CSF. As the purpose of the work is not the study of the HVS but its applications in a video coding framework, the psychophysical experiments have been performed to study perception of coding noise. This has been done in the following way: five subjects took part to experiments where they have been asked to assess the visibility of stimuli. White noise filtered by a perceptual channel has been used as stimuli as it will represent a signal close to coding noise filtered by the same channel. The experiment was a two alternatives forced choice discrimination task [10]. The level of the stimuli were adaptively decided on the fly by a modified PEST procedure [11]. Details of the experiment are reported in [12].

An example of measurements is presented in Fig. 5. The

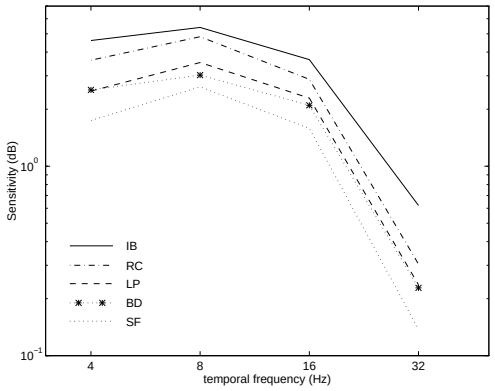


**Figure 5. Graph of the measured sensitivity for the five subjects as a function of temporal frequency and at a spatial frequency of 4 cpd.**

temporal sensitivity measured at a spatial frequency of 4 cpd has been plotted for the five subjects. Each data point is the average of three successful measurements. The sensitivity of the subjects varies but the general shape of the curve is very consistent with the theoretical prediction [4]. Further measurement, reported in [12] permitted the estimation of the whole CSF. A contour plot of the estimated curve is shown in Fig. 6, as a function of spatial and temporal frequency.

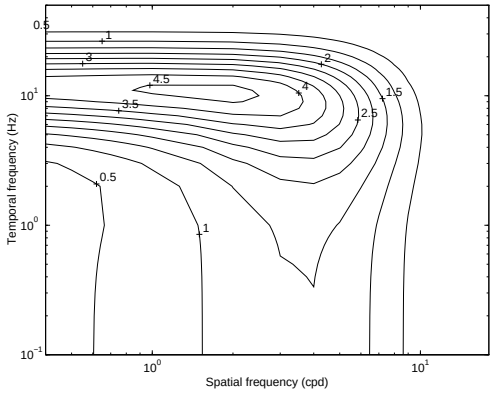


**Figure 6. Contour plot of the estimated spatio-temporal CSF.**

## 6. EXAMPLE OF APPLICATION

A typical application of such a model is objective quality assessment. A perceptual metric for the assessment of video coding quality has been designed and is described in [13]. Basically, the metric works as follows: the architecture presented in Sec. 4. is used. The original and error sequences are decomposed using the Gabor filter bank. Contrast sensitivity and masking are used to predict the perceived error that is then expressed in jnd's.

This signal is then divided into three-dimensional blocks. The block dimensions are chosen as follows: the temporal

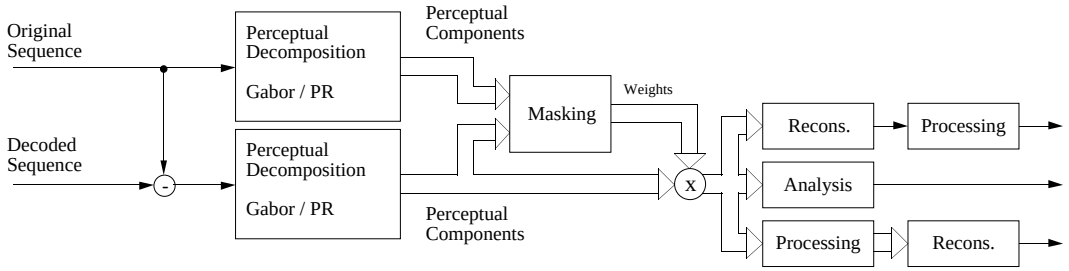


**Figure 4. The general structure of a perceptual processing of video sequences. The thick arrows represent a set of perceptual components. The thin lines represent sequences.**

dimension is chosen to account for persistence of the images on the retina. The spatial dimension is chosen to consider focus of attention, i.e. the size is computed so that a block covers two degrees of visual angle, which is the dimension of the fovea. The data is then pooled, for each block, over all channels by probability summation [4]. This yields a distortion measure for each block. A global measure for the whole sequence can then be obtained by averaging this measure over blocks.

The metric, denoted Moving Pictures Quality Metric (MPQM) has been used to characterize the subjective quality of MPEG-2 [14] coding performance over a range of bitrates. It turned out that the metric correlates quite well with other subjective evaluations of MPEG-2 [15].

## 7. CONCLUSION

This paper presented a spatio-temporal model of human vision. It models the multi-channel structure of the primary visual cortex, contrast sensitivity, masking and spatio-temporal interaction in human vision. The model has been parameterized for a video coding framework by the means of psychophysics. The resulting scheme predicts the response of the neurons of the primary visual cortex and can be used as a general architecture for perceptual processing of video sequences. A quality metric for coded video sequences has been built on top of the model and is briefly described. The tool proved to correlate well with subjective data.

## REFERENCES

[1] R. L. De Valois and K. K. De Valois. *Spatial Vision*. Oxford University Press, 1988.

[2] J. G. Daugman. "Spatial Visual Channels in the Fourier Plane". *Vision Research*, Vol. 24, pp. 891–910, 1984.

[3] J. G. Daugman. "Two-Dimensional Spectral Analysis of the of Cortical Receptive Field Profiles". *Vision Research*, Vol. 20, pp. 847–856, 1980.

[4] Andrew B. Watson. *Handbook of Perception and Human Performance*, Vol. 1, Sensory Processes and Perception, Chapter 6, Temporal Sensitivity. John Wiley, 1986.

[5] S. T. Hammett and A. T. Smith. "Two Temporal Channels or Three? A Re-evaluation". *Vision Research*, Vol. 32, No. 2, pp. 285–291, 1992.

[6] A. B. Watson and J. A. Solomon. "Contrast Gain Control Model Fits Masking Data". on http://vision.arc.nasa.gov, 1995.

[7] Patrick C. Teo and David J. Heeger. "Perceptual Image Distortion". In *Proceedings of the International Conference on Image Processing*, pp. 982–986, Austin, TX, November 1994.

[8] Christina A. Burbek and D. H. Kelly. "Spatiotemporal Characteristics of Visual Mechanisms: Excitatory-Inhibitory Model". *Journal of the Optical Society of America*, Vol. 70, No. 9, pp. 1121–1126, September 1980.

[9] Serge Comes. *Les traitements perceptifs d'images numerisées*. PhD thesis, Université Catholique de Louvain, 1995.

[10] J. C. Falmagne. *Handbook of Perception and Human Performance*, Vol. 1, Sensory Processes and Perception, Chapter 1, Psychophysical Measurement and Theory. John Wiley, 1986.

[11] J. L. Hall. "Hybrid Adaptive Procedure for Estimation of Psychometric Functions". *Journal of the Acoustical Society of America*, Vol. 69, No. 6, pp. 1763–1769, June 1981.

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

[13] 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 IS&T Symposium on Electronic Imaging: Science and Technology*, San Jose, CA, January 28 - February 2 1996. IS&T/SPIE. accepted for publication.

[14] Draft ISO-IEC/JTC1/SC29/WG11, Motion Picture Expert Group. "MPEG–II: Test model 4, January 1993.

[15] Andrea Basso, İsmail Dalgıç, Fouad A. Tobagi, and Christian J. van den Branden Lambrecht. "Study of MPEG-2 Coding Performance based on a Perceptual Quality Metric". In *Proceedings of the Picture Coding Symposium*, Melbourne, Australia, 1996. accepted for publication.