Objective Video Quality Assessment by Image Comparison*

P. Brétillon[1], S. Olsson[2], J. Baïna[1].

[1] TDF-C2R France, [2] TERACOM AB Sweden.

jbaina@c2r.tdf.fr

Abstract

The concern of quality of service in distribution networks for MPEG2 digital television has contributed to increase the need for new objective video quality assessment methods. The bitrate reduction techniques employed and transmission errors can generate specific impairments. In the context of a real digital TV service, their impact on the video quality perceived by end-users has to be evaluated. In this perspective, a direct comparison between a degraded and a reference sequence is often undertaken. The main objective of the paper is to provide the background knowledge necessary to understand the difficulties of the field. This paper also gives an overview of the algorithmic possibilities that exist and the progress that has been made so far. Several techniques are presented.

1. INTRODUCTION

The advent of digital TV implies a new path for the signals, which impact on quality is different from analog TV. A major reason relies in the non linear and content-adaptive processing of the signals. The digital representation of the compressed and transmitted images is also responsible for an abrupt quality degradation. This paper concerns video quality which is one of the most pertinent component in the quality of MPEG2 TV service.

The most reliable tool for quality assessment is still subjective testing with a panel of viewers. However, the cost of such processes in terms of time, manpower and required experience is very high. Furthermore, digital image processing have led to increase of the amount of processing equipments with a variable impacts on the final quality of the pictures. Objective measurements could solve, at low cost, the problems linked to the assessment of service quality. In addition, objective measurements would allow a continuous picture quality monitoring of digital encoding and broadcasting equipment in operative service, as well as a practical comparison tool for coders. However, the performance of such algorithms in terms of approximation of the quality perceived by the viewer must be high enough. Perceived quality is difficult to evaluate directly, because many factors must be taken into consideration.

There are several approaches to video quality evaluation. In this paper, we mainly present those that are based on perceptual models. A few other methods that rely on the error signal are also briefly presented, along with their major advantages and drawbacks. The main

* This work is supported by the European Commission in the framework of the ACTS QUOVADIS project involving 11 different partners : TDF-C2R (Prime Contractor), EBU, IRT, Rohde & Schwarz, Retevision, Matra Communication, FUB, RAI, TERACOM, TELMAT, CCETT.

The global budget of the project is 6.9 MECU with a 2.1 MECU contribution from the EC.

elements of perceptual model based methods are then presented, along with the underlying properties of the Human Visual System used. We conclude on the performances and applicability of this approach in the context of quality of service in digital television networks.

2. QUALITY ASSESSMENT METHODS

The impairments introduced in a video sequence by a processing system can be considered as an additive error. The error sequence is computed on a pixel by pixel and frame by frame basis and is an important tool frequently used to analyze a coder. Its amplitude and statistical properties carry information about the characteristics of the distortions generated by the video system. Several approaches to use it as the base of an objective quality assessment are exposed in the following.

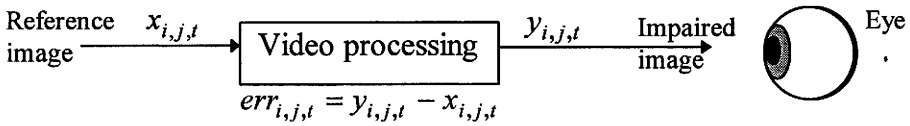

Figure 1: The error signal is a basic tool for digital system analysis.

2.1 Error signal analysis

This approach relies on the a priori knowledge of the coding algorithms. Taking advantage of it, the generated impairments characteristics can be known. This allows an efficient analysis of the error signal: for each impairment characteristic, a feature can be processed, eventually taking some properties of the visual system into account. The final set of features is usually summed by an arbitrary model to get an objective image quality. The model parameters are adjusted by a step of optimization of the correlation with subjective notes on a set of pictures [1].

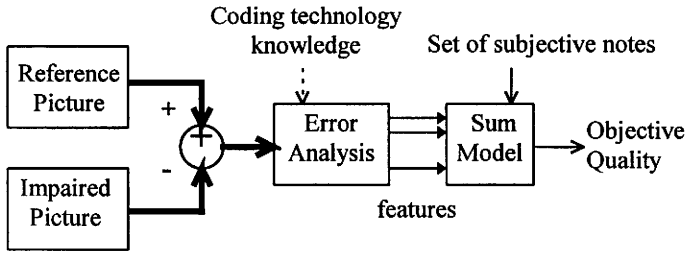

Figure 2: Quality evaluation by error analysis

A typical example is the block DCT based coder like JPEG or MPEG. At low bit rates, this kind of system produces a very noticeable mosaic pattern called blocking effect, that reveals the underlying block structure. An efficient feature analyses the error image in the vicinity of block borders to reveal this impairment. This kind of approach is efficient because it tracks specific impairments. It has mainly been applied to JPEG still images, with a good correlation of 0.88, but rarely to video [2]. However, the correlation drops for images which are not in the initial set. The strong link of the method with the compression technology used is another limitation.

2.2 General approaches

Signal to Noise Ratio (SNR) and its derivatives (PSNR) are widely used to analyze analog systems. However, their application to digital systems is usually not satisfactory, at least in terms of a good quality evaluation. Digital systems like an MPEG codec are not linear. The error can not be any more considered as an independent and stationary random noise. It is correlated to the signal content and depends on the processing algorithms too. Secondly, these parameters cannot discriminate between a few high amplitude errors (annoying for final viewers) and many low amplitude impairments (maybe subjectively imperceptible, or less annoyable). Consequently, the SNR, which is based on the error power, is not accurate enough for quality assessment.

To overcome the limitations of SNR, several adaptations have been proposed and tested. They take into account the frequency dependent visibility of noise. Several spatial frequency weighting functions have been measured or normalized [3,4]. There is also the fact that human quality judgment seems to be primarily based on the most visible errors. Instead of a global SNR, local SNR are evaluated on independent sub-image, typically blocks. The biggest errors are taken into account with a bigger weight to evaluate the final quality. These methods provide a better correlation with subjective tests, but are generally not satisfactory. To improve the correlation, more characteristics of the eye must be integrated.

The visibility of a single stimulus in flat areas is better than in areas including edges. This effect is called spatial masking, but it also exists for temporal frequencies. The integration of visual spatio-temporal frequency response and visual masking effect into a single process leads to a modified 3D-SNR [5]. The approximate HVS temporal frequency characteristic W is computed as the product of the eye spatial and temporal frequencies characteristics. The relationship between the entropy of a block and a masking coefficient has been found such that the masking effect is taken into account in the evaluation. The final quantity 3D-SNR is computed using a noise measure which has been masked using the masking coefficient and weighted with W.

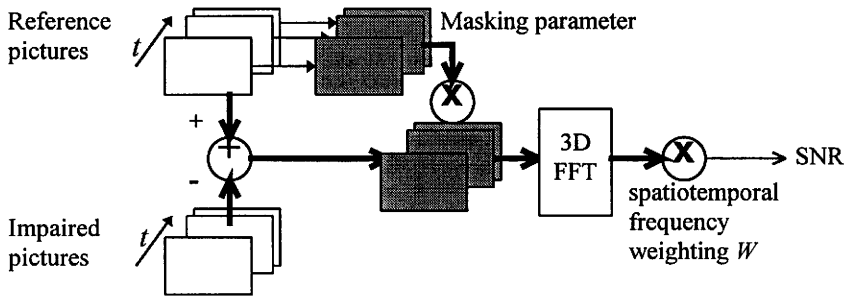

Figure 3: Weighted 3D-SNR [5].

2.3 Perceptual model based approaches

This approach is motivated by the fact that the Human Visual System (HVS) is not linear. Stimuli of the same amplitude, but different spatial and temporal frequencies are not perceived in the same way by the human eye. Furthermore, the perception is different when a stimulus is included in flat spatial areas or in areas including edges or details. Knowledge of the HVS model would have profitable effects when applied, in automatic quality evaluation,

as well as in the improvement of encoding systems. Several properties have already been used into account in the methods above. The HVS model allows to map the studied signals to a "perceptual" domain, in which the measured differences are expected to be close to the visual effect they produce. In this space, the amplitude is usually called contrast.

In this part, some additional characteristics are integrated into the model. The main one is a multichannel decomposition structure. This has been applied to a single picture quality evaluation [6,7]. Some research activities of van den Branden and Westen *et al.* aim at developing metrics based on these concepts [8,9]. Figure 4 shows the main elements of these metrics, based on the HVS model further presented in section 3.

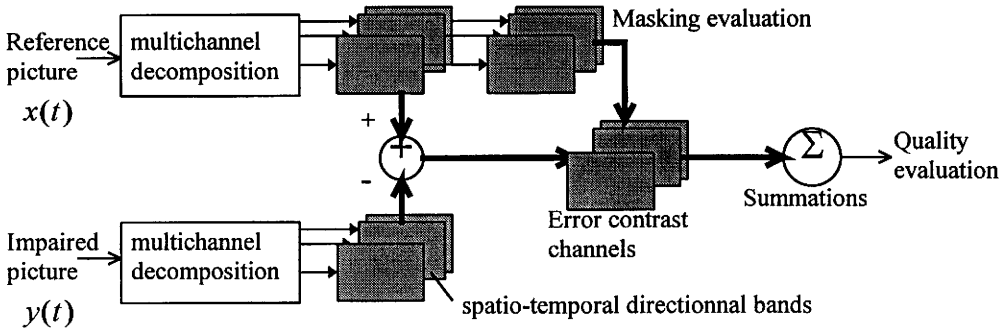

Figure 4: Quality evaluation with HVS model

A spatio-temporal model of human vision has been developed for the assessment of video coding quality [8,9]. A three-dimensional filter bank simulates the multichannel characteristic of the HVS according to 5 spatial frequencies, 4 orientations and 2 temporal frequencies. A distortion measure is computed, accounting for the higher levels of cognition in the brain. At this stage, the metric also accounts for the focus of attention and is computed over spatio-temporal blocks of the sequence. Their dimensions are chosen in agreement with persistence of the images on the retina, and the field size of detailed vision (two degrees of visual angle). The distortion measure is computed for each block by pooling the error over the channels, and finally mapped onto a quality scale from 1 to 5, called MPQM (Moving Picture Quality Metric). The metric has been found to be able to detect the saturation in quality that occurs at high bitrates according to subjective quality assessments. At lower bitrates, the metric also exhibits a behavior that matches correctly human judgment.

3. HUMAN VISUAL SYSTEM MODEL ELEMENTS

The HVS has been extensively studied for its sensibility to a stimulus (here, the error image) on a background (the actual image). Thus the use of these results is straightforward in principle. However, some factors limit the derivation of a quality assessment method. First, the studied stimuli are very simple, and can't account for the variety of natural images. Secondly, the results are given as a "just noticeable difference" threshold of visibility (JND), whereas real degradation can be far from this situation. Specific adaptations may be used.

3.1 Amplitude response

The most well known property of the HVS is its amplitude non linearity. The Weber-Fechner law defines the minimum perceptible intensity difference, and shows that the response to amplitude is approximately logarithmic, except in the dark and very bright ranges [11].

The behavior for colored light is similar, but the eye has less sensibility for it.

3.2 Spatial or temporal frequencies sensibility

The simplest hypothesis which has been used to explain some experimental data is to approximate the human visual system with a linear filter. In order to estimate the transfer function $H(\omega)$ of this filter, a solution which is widely adopted is to measure the Contrast Sensitivity Function (CSF). At each frequency ω, the minimum contrast $C(\omega)$ necessary to distinguish the sinusoidal gratings from a uniform background is measured. This contrast is called the contrast threshold. The contrast sensitivity function is then defined by

$$H(\omega) = CSF(\omega) = 1 / C(\omega) \tag{1}$$

Figure 5 shows that the contrast sensitivity is high at low spatial frequencies. It is maximum around 5 cycles/degree and decreases rapidly for high spatial frequencies. However, the precise meaning of high and low frequency varies with the wavelength of incident light, because the quality of the retinal image varies strongly with wavelength. There is less sensitivity for blue than for red or luminance. Spatial frequency sensitivity is not equal for all directions. It is considered to be equal for horizontal and vertical gratings, but about 1.4 times smaller for diagonal directions [12].

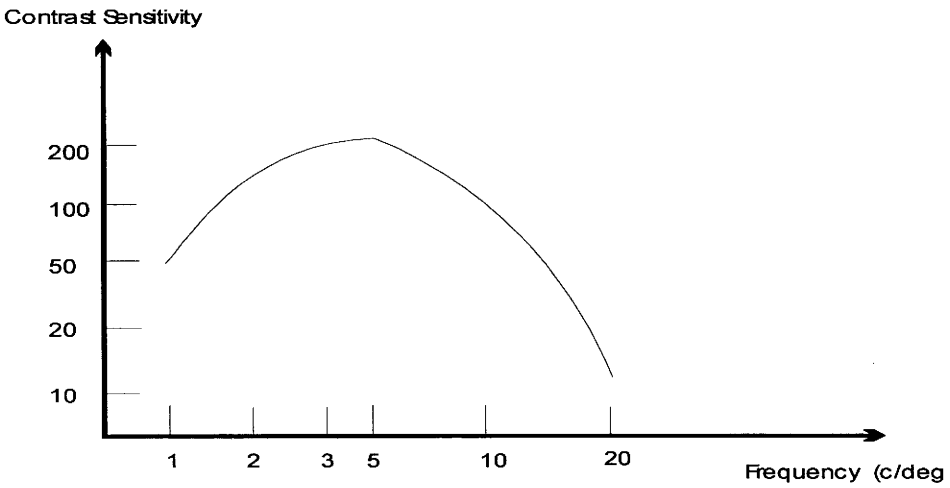

Figure 5: Contrast Sensitivity Function

The temporal contrast sensitivity can be measured in a similar way. It has the same shape than for spatial frequencies. However, spatial and temporal characteristics are not separable. A temporal frequency greater than 6Hz will result in a spatial resolution loss, and conversely a spatial stimulus greater than 6 cycles/degree will result in a temporal resolution loss (provided the eye does not move to track the moving object).

3.3 Multichannel models

Several experiments have shown that the retinal image is likely to be processed in separate frequency channels [13]. The assumption is that the retina decomposes an image through independent bandpass linear filters. It has been shown that these filters have approximately the same bandwidth on a logarithmic scale, and have a spatial orientation selectivity, with an angular resolution of about ±20°. This concept has been extended to temporal channels, which are tuned for direction of motion. The multichannel model is illustrated in Figure 6. The contrast associated to a channel is defined with respect to the lower frequencies channels. Extensive multiresolution models can be found in the literature [14,15,16].

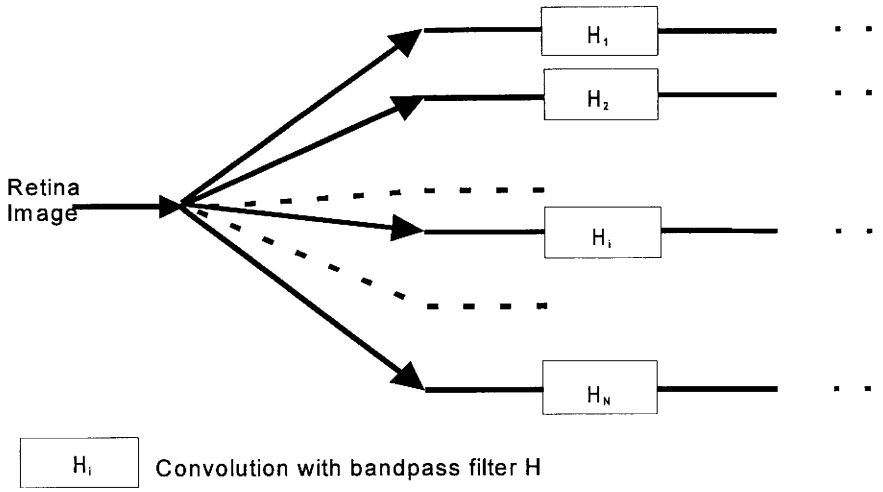

Figure 6: Multichannel model: independent bandpass filters

Using this model to evaluate the quality of natural images requires to take the interactions between neighboring channels into account. This phenomenon is called masking. Both spatial and temporal masking basically imply that impairments are less visible in the vicinity of a spatial or temporal discontinuity of high contrast. For example, the presence of a temporal change in the sequence, like a scene cut, makes low-amplitude details less visible for the next few TV frames, implying that spatial detail and motion can be masked immediately before and after the cut. Low-contrast discontinuities, on the contrary, can lead to a higher visibility of impairments on the bright side of the discontinuity, i.e., the presence of an edge or a temporal jump can make impairments more visible. A final effect that can be regarded as masking is the decreased visibility of impairments at very low and very high gray values compared to the visibility at medium gray. [15,16,17]

The masking function can be modeled by an increase of the threshold of visibility. Several models can be found in [6,17,18]

3.4 Impairment summation

The sensibility of the HVS has been evaluated in terms of Just Noticeable Difference threshold. From the point of view of quality, it is preferable to have a progressive value, like a probability of detection, as the amplitude of the impairment increases. This is achieved by the psychometric function.

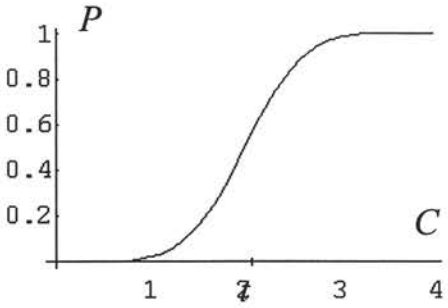

Figure 7: Sample plot for a psychometric function

The psychometric function in Figure 7 gives the probability of detection P with respect to the contrast C of the difference stimulus. The threshold t takes into account both HVS frequency sensibility and masking [19].

The introduction of a probability of detection allows to sum the probabilities between the channels, which gives a single map of probability of detection. Finally, a single number for the objective quality of a picture or a short video sequence can be computed by another summation on the map. However, this last step involves higher levels of cognition in the brain, like the notion of focus of attention. The eye can accurately view only a field of 2 by 2 degrees, and it has been observed that some zones are more attractive than others for an average viewer. Thus the spatial location of the degradation is important to evaluate quality. There is ongoing research on this subject [20].

4. Conclusion

In this paper, several approaches of objective picture or video sequence quality assessment based on an error signal have been presented. The integration of eye properties into the assessment method allows to predict video quality better. However, some limitations are found in the case of strong degradation. Ongoing research on Human Visual System should bring improvements in this field. The methods presented in this paper are well adapted to applications such as video coder quality evaluation, as a sample test as well as continuous monitoring. Their use in broadcast quality monitoring is possible too, but then the condition of reference availability is difficult to fulfill. We feel that an interesting case would be satellite broadcasting.

References

[1] I. Davies *et al.*, "Automated image quality assessment", SPIE Vol. 1913 Human Vision, Visual Processing, and Digital Display, 1993, pp 27-36

[2] Algazi *et.al.*, "Important Distortion Factors in the Encoding of Very High Quality Images", SPIE Vol. 2298, *Proc. Human Vision, Visual Processing and Digital Display V,* San Diego (CA), USA, March 1994.

[3] "Transmission performance of television circuits designed for use in international connections", Recommendation ITU-R 567-3, Annex II, Part C, §3

[4] P. Barten, "Evaluation of subjective image quality with the square root integral method", J. Opt. Soc. Am. A, Vol. 7 No. 10, Oct 1990

[5] J. Okamoto *et al.*, "A study on subjective and objective evaluation methods for coded moving picture quality", International Picture Coding Symposium, Melbourne, Australia 13-15 March 1996, pp.519-523

[6] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity", Proceedings of SPIE, Vol.1616, pp.2-15, 1992

[7] P. Teo and D. Heeger, "Perceptual image distortion", Proceedings of the International Conference on Image Processing, pp.982-986, Austin, TX, November 1994

[8] C. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image restoration and quality assessment applications", Proceedings of the ICASSP 1996, submitted paper

[9] S. Western *et al.*, "Perceptual image quality based on a multiple channel HVS model", ICASSP, pp.2351-2354, 1995

[10] A. Basso *et al.*, "Study of MPEG-2 coding performance based on a perceptual quality metric", International Picture Coding Symposium, Melbourne, Australia 13-15 March 1996, pp.263-268

[11] A. K. Jain, "Fundamentals of digital image processing", Prentice Hall, 1989.

[12] W.E. Glenn, K.G. Glenn, C.J. Bastian "Imaging system design based on psychophysical data", (New York Institute of technology), Proceedings of the SID, Vol. 26/1, 1985, pp 71-78

[13] F. Campbell and J. Robson, "Application of Fourier analysis to the visibility of gratings", J. Physiology, Vol.197, pp. 551-566, 1968

[14] A. Watson, "Perceptual-component architecture for digital video", J. of the Optical Society of America, Vol. 7, No.10, pp. 1943-1954, October 1990

[15] A. Netravali, "Picture coding: A review", Proceedings of the IEEE, Vol.68, No.3, March 1980, pp.366-406

[16] B. Wandell, Foundations of Vision: Behaviour, Neuroscience, and Computation", Draft of December 25, 1994

[17] B. Girod, "The information theoretical significance of spatial and temporal masking in video signals", SPIE Vol. 1077 Human Vision, Visual Processing, and Digital Display (1989), pp.178-187

[18] P. Barten, "Simple model for spatial frequency masking and contrast discrimination", SPIE Vol 2411 (Human Vision, Visual Processing, and Digital Display VI), Feb. 1995 (modelsvh.doc)

[19] Peter G.J. Barten, "Spatio-temporal model for the contrast sensitivity of the human eye and its temporal aspects", SPIE Vol 1913 (Human Vision, Visual Processing, and Digital Display IV), Feb. 1993.

[20] M. Ardito, M. Visca, "Correlation between objective and subjective measurements for video compressed systems", IBC Convention, Amsterdam, September 1995.