Study of MPEG-2 Coding Performance based on a Perceptual Quality Metric*

Andrea Basso* ** İsmail Dalgıç* Fouad A. Tobagi* and
Christian J. van den Branden Lambrecht***

** Department of Electrical Engineering Stanford University*
*** Telecommunication Laboratory Swiss Federal Institute of Technology*
**** Signal Processing Laboratory Swiss Federal Institute of Technology*

Abstract

In this paper some well known quality metrics such as PSNR and the metric developed at Institute for Telecommunication Sciences (ITS) are reviewed. Their shortcomings in measuring quality of coded video compared to subjective tests are pointed out. Then, a new video quality metric called Moving Picture Quality Metric (MPQM) is presented. This metric models the human visual system and matches correctly subjective evaluations. Comparative results in the case of constant bit rate (CBR) MPEG-2 coded sequences are presented, showing the superiority of MPQM over ITS and PSNR.

1 Introduction

The interest in multi-media applications - with a strong emphasis on video issues - is growing tremendously. Video assessment is a fundamental and still not sufficiently explored aspect of the current research on video coding. Visual objective metrics that are coherent with quality as perceived by human observers are beginning to emerge only recently. Furthermore very recently the concept of constant-quality video encoding has been introduced in [1] and further developed in [2].

The motivation of this work is to determine the right kind of quality metric for devising a constant-quality variable bit rate (CQ-VBR) video encoding scheme for MPEG-2 in view of an evaluation its performances over ATM.

In this paper some well known quality metrics for video such as the well known and widely used Peak Signal to Noise Ratio (PSNR) and the metric developed at Institute for Telecommunication Science (ITS) are reviewed. Then, a new video quality metric called Moving Picture Quality Metric (MPQM) is discussed. That metric is based on a model of the early stages of the human visual system and it matches subjective evaluations correctly. Results for the considered metrics are shown for constant bit rate (CBR) MPEG-2 coded sequences.

The paper is organized as follows. Sec. 2 overviews the literature on video quality assessment and measurement. Sec. 3 illustrates the inadequacy of PSNR metric for video quality assessment by means of a simple example. Sec. 4 presents the ITS video quality metric and its performance on MPEG-2 coded video. Sec. 5 discusses the MPQM. Some conclusive remarks end the paper.

2 State of the art in video quality assessment and measurement

2.1 First Developments

Video quality assessment plays a fundamental role in the development of new and existing video coding algorithms. The interest for image quality evaluation has been strong since the sixties. Several image quality metrics have been developed, such as the Strehl measure [3], based on the degradation to which an image is subject to, whenever it passes different real optical systems compared to the ideal case. Attempts with bivariate metrics of image quality have been done in particular during the seventies. The reader is referred to [4] for a review.

The mean square error has been retained for its property of being easily analytically tractable. Wilder [5] did a rather complete evaluation of dif-

*This work was supported in part by NSF under grant NCR-9016032, and by Pacific Bell.

ferent mean-square error metrics including power-law, logarithmic, gradient and Laplacian transformations.

More recently, with the advent of the digital television and the development of new video coding standards such as H.261, MPEG-1 and MPEG-2 the need for effective perceptual quality metrics has became even more stringent. International committees, such as CCIR and MPEG, have devoted a lot of effort in the study of video quality assessment [6, 7, 8, 9].

We can summarize the recent efforts in image and video quality assessment in two different lines of thought.

2.2 Metrics Based on Subjective Tests

Such metrics are based on linear or non-linear combinations of distortion measures applied to features extracted from images and video. Such features include false edge effect, blocking effect, blurred edges effect and temporal distortions. The coefficients of the linear or non-linear combinations are chosen in order to maximize the correlation with subjective tests.

Pioneer on this area is Miyahara [10] with a method that extracts five features and on this basis computes a simple model which is then tuned to maximize correlation with subjective tests. On the same line F. Lin *et al.* [11], Davies [12] and Webster [13] have developed a non-linear neural network model for image quality assessment. On the same principles a quantitative video quality metric has been designed by Wolf *et al.* [14] at Institute for Telecommunication Sciences (ITS) in Colorado.

2.3 Metrics based on Human Visual System

The second line of thought bases the metric on physiological evidences. Following this concept the metric mimics the human evaluation by modeling closely the initial stages of the human visual system.

Several physiological and psychophysical experiments have been carried out on perception. Measures based on physiological evidences have been conducted by Mannos and Sakrison, with particular focus on mean square error quality criterions for monochrome images on the basis of a simple model of the human visual system. These considerations are discussed in the Rec.500-3 from CCIR [6] on subjective assessment of the quality of television pictures.

De Valois [15] conducted electro-physiological experiments showing that the cells of the primary visual cortex are tuned to bands in spatial frequency and orientation. Physiological experiments conducted by Daugman [16] confirmed the presence of several mechanisms of vision that divide the frequency plane into frequency and orientation bands.

Several models have been developed which make use of masking, contrast sensitivity, multichannel structure [17, 18, 19].

The research activities of van den Branden [20] Comes [21] and Western *et al.* [22] are the first efforts in the development of metrics based on these concepts.

2.4 Subjective evaluation of CBR MPEG-2

In [9], results of a subjective evaluation for CBR MPEG-2 sequences encoded at various rates from 4 to 30 Mb/s are given. These results indicate that from 4 to 9 Mb/s, there is a significant increase in subjective quality, and the quality degradations become nearly undetectable at rates greater than 15 Mb/s. Therefore, an appropriate quantitative video quality metric has to be able to capture the degradations in the 4-15 Mb/s range, and should indicate near-perfect quality beyond that range.

In the following sections the well known PSNR, the metric developed at ITS and a new metric based on this latter approach will be discussed.

3 PSNR as video quality metric

It is well known that PSNR does not correlate well with subjective video quality assessments [23]. Peak Signal to Noise Ratio is defined as:

$$PSNR = 10 \log_{10} \frac{255^2}{\sigma^2} \qquad (1)$$

The value 255 represents the peak value that each pixel can have (assuming 8 bit resolution). The symbol σ^2 is defined as:

$$\sigma^2 = \frac{1}{N} \sum_{i=1,N} (o_i - r_i)^2 \qquad (2)$$

where o_i represents the ith pixel in the original image while r_i represents the ith pixel in the distorted image.

We illustrate the inadequacy of PSNR measure with the following simple example. We applied the PSNR measure to CBR MPEG-2 coded sequences at bitrates ranging from 3 Mb/s to 30 Mb/s. For all the rates considered, the buffer size was kept to its maximum value to minimize any influence on the image quality. All the other coding parameters were kept identical. In Fig. 1 we show the average PSNR for *Basketball* test sequence CBR and MPEG-2 coded as a function of bit rate, in the range 3-30 Mb/s. While from a subjective quality point of view the user is not able to make any distinction between the original and the coded sequence in the range 15-30 Mb/s, the PSNR measure is indicating a large variation, more than $6dB$, in image quality. The large discrepancy between the subjective quality evaluation and the assessment given by the PSNR metric makes it an unsuitable metric for video.

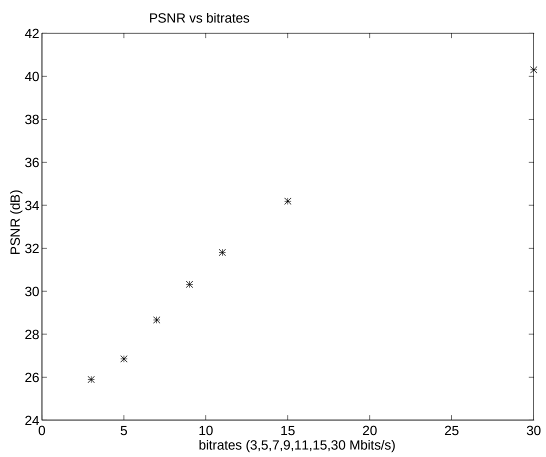

Figure 1: Average PSNR for *Basketball* test sequence CBR and MPEG-2 coded as a function of bit rate

4 ITS Quantitative Video Quality Metric

In this section we will briefly illustrate one of the most commonly used quality measure and the only one available. It has been developed at *Institute for Telecommunication Science* in Colorado. We will show that this measure in inadequate for evaluating the quality of MPEG-2 coding schemes.

To design this measure, the authors first conducted a set of subjective tests in accordance with CCIR Recommendation 500-3 [6], which specifies viewing conditions, rating scales, etc. The viewers were shown a number of original and degraded video pairs, each of them 9 seconds long, and they were asked to rate the difference between the original video and degraded video as either imperceptible (5), perceptible but not annoying (4), slightly annoying (3), annoying (2), or very annoying (1).

As described in [13], the quantitative measure is a linear combination of three quality impairment measures. Those three measures were selected among a number of candidates such that their combination matched best the subjective evaluations. The correlation coefficient between the estimated scores and the subjective scores was 0.94, indicating that there is a good fit between the estimated and the subjective scores. The standard deviation of the error between the estimated scores and the subjective scores was 0.4 impairment units on a scale of 1 to 5; thus, differences below 0.4 should not be considered significant.

The quantitative measure is based upon two quantities, namely, *spatial information* (SI) and *temporal information* (TI). The spatial information for a frame F_n is defined as

$$SI(F_n) = STD_{space}\{Sobel[F_n]\},$$

where STD_{space} is the standard deviation operator over the horizontal and vertical spatial dimensions in a frame, and $Sobel$ is the Sobel filtering operator, which is a high pass filter used for edge detection.

The temporal information is based upon the motion difference image, ΔF_n, which is composed of the differences between pixel values at the same location in space but at successive frames (i.e., $\Delta F_n = F_n - F_{n-1}$). The temporal information is given by

$$TI[F_n] = STD_{space}[\Delta F_n].$$

Note that SI and TI are defined on a frame by frame basis. To obtain a single scalar quality estimate for each video sequence, SI and TI values are then time-collapsed as follows. Three measures, m_1, m_2, and m_3, are defined, which are to be linearly combined to get the final quality measure. Measure m_1 is a measure of spatial distortion, and is obtained from the SI features of the original and degraded video. The equation for m_1 is given by

$$m_1 = RMS_{time}(5.81\left|\frac{SI[O_n] - SI[D_n]}{SI[O_n]}\right|),$$

where O_n is the n^{th} frame of the original video sequence, D_n is the n^{th} frame of the degraded video sequence, and RMS denotes the root mean square function, and the subscript *time* denotes that the function is performed over time, for the duration of each test sequence.

Measures m_2 and m_3 are both measures of temporal distortion. Measure m_2 is given by

$$m_2 = f_{time}[0.108MAX\{(TI[O_n] - TI[D_n]), 0\}],$$

where

$$f_{time}[x_t] = STD_{time}\{CONV(x_t, [-1, 2, -1])\},$$

STD_{time} is the standard deviation across time (again, for the duration of each test sequence), and $CONV$ is the convolution operator. The m_2 measure is non-zero only when the degraded video has lost motion energy with respect to the original video.

Measure m_3 is given by

$$m_3 = MAX_{time}\{4.23LOG_{10}(\frac{TI[D_n]}{TI[O_n]})\},$$

where MAX_{time} returns the maximum value of the time history for each test sequence. This measure selects the video frame that has the largest added motion. This may be the point of maximum jerky motion or the point where there are the worst uncorrected errors.

Finally, the quality measure $\hat{s}$ is given in terms of m_1, m_2, and m_3 by

$$\hat{s} = 4.77 - 0.992m_1 - 0.272m_2 - 0.356m_3.$$

In Figure 2 we show $\hat{s}$ as a function of the bitrate under the same evaluation conditions as in the PSNR case. As shown in the figure, the ITS metric is able to capture the saturation of the perceived quality in the range 15-30 Mb/s; thus it is an improvement with respect to PSNR. On the other hand, the variation of the metric in the 3 Mb/s to 15 Mb/s range is only 0.2 impairment units, the absolute value of the metric being 4.5 even at 3 Mb/s. Therefore, the degradations in the 3-15 Mb/s range are not captured. The reasons for that are twofold. First of all, the metric has been designed for low bitrate sequences, which have significantly more coding artifacts compared to typical MPEG-2 encoded sequences. Secondly, the metric is not always capable of capturing correctly the DCT coding artifacts such as blocking effects and mosquito noise. Thus, this metric is not suitable for quality assessment of MPEG-2 encoded video sequences.

5 Moving Pictures Quality Metric (MPQM)

The previous discussion illustrates the shortcoming of the existing image quality metrics. The reason is that human vision is a very complex process and those metrics are not able to match its behavior. Several studies have shown that a correct estimation of subjective quality has to incorporate some modeling of human vision [23, 24, 21].

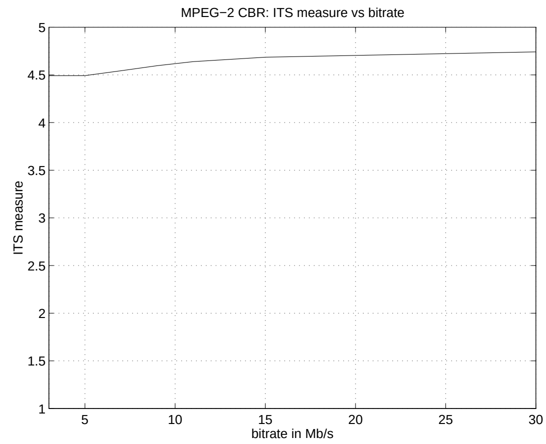

Figure 2: $\hat{s}$ for *Basketball* test sequence CBR and MPEG-2 coded as a function of bit rate.

A spatio-temporal model of human vision has been developed for the assessment of video coding quality [20, 25]. The model is based on the following properties of human vision:

- The responses of the neurons of the primary visual cortex (called area V1) are bandlimited. The human visual system has a collection of mechanisms or detectors (called channels) that mediate perception. A channel is characterized by a localization in spatial frequency, spatial orientation and temporal frequency. Such channels are simulated by a three-dimensional filter bank.

- In a first approximation, those channels can be considered to be independent. Perception can thus be assessed channel by channel without interactions.

- Perception in a channel is characterized by two phenomena: contrast sensitivity and masking. Human sensitivity to contrast is known to be a function of frequency (spatial and temporal) as well as orientation. This leads to the concept of *contrast sensitivity function*, which specifies the threshold of detection for a stimulus as a function of frequency. Masking accounts for inter-stimuli interferences. It is known that the presence of a background stimulus will modify the detection of a foreground stimulus. Masking thus corresponds to a modification of the detection threshold according to the contrast of the background.

The working model described in [20] incorporates the above described considerations of the HVS. The filter bank used in the model decomposes the data according to 5 spatial frequencies, 4 orientations and 2 temporal frequencies. It has been especially parameterized for the framework of video coding by means of psychophysical tests.

Such a model permits to predict the response from the neuron in area V1 and thus the *perceived* distortion. This is done by a decomposing the data in perceptual channels and predicting perceived stimuli using contrast sensitivity and masking. Thereafter, a distortion measure is computed, accounting for the higher levels of cognition of the brain. At this stage, the metric also accounts for the focus of attention and is computed over blocks of the sequence. Such blocks are three-dimensional and their dimensions are chosen as follows: the temporal 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 distortion measure is computed for each block by pooling the error over the channels. Basically, the magnitude of the channels' output are combined by Minkowski summation with a higher exponent to weight the higher distortions more. The actual computation of the distortion M_E for a given block is computed according to Eq. 3:

$$M_E = \left(\frac{1}{N} \sum_{c=1}^{N} \left(\frac{1}{N_q} \sum_{t=1}^{N_t} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} |e[x,y,t,c]| \right)^{\beta} \right)^{\frac{1}{\beta}}, \tag{3}$$

where $e[x,y,t,c]$ is the masked error signal at position (x,y) and time t in the current block and in the channel c; N_x, N_y and N_t are the horizontal and vertical dimensions of the blocks; N is the number of channels and $N_q = \frac{1}{N_x N_y N_t}$. The exponent of the Minkowski summation is β and has a value of 4, which is close to probability summation [26].

In this application, the error measure M_E is further mapped onto a quality scale from 1 to 5 according to the following function, relating the error measure to the quality index MPQM:

$$\text{MPQM} = \frac{5}{1 + N M_E},$$

where N ensures a mapping between 1 and 5. This free parameter has been estimated on the basis of the vision model [25] and has a value of $N = 0.623$.

The previous experiment has been repeated under the same conditions with MPQM. Results are depicted in Fig. 3. It can be seen that the metric is able detect the saturation in quality that occurs at high bit rates according to the subjective quality assessment. At lower bit rates, the metric exhibits a behavior that matches correctly human judgment [9].

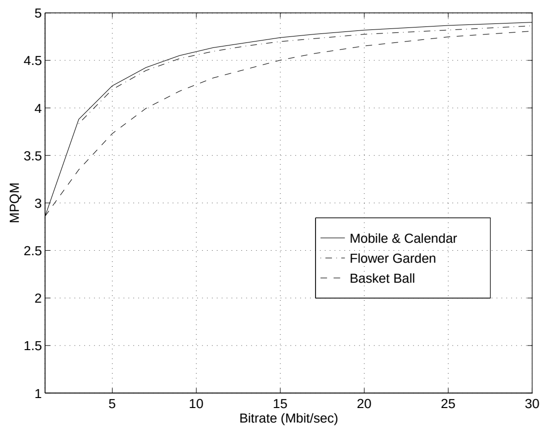

Figure 3: MPQM as a function of the bitrate for MPEG-2 encoding of the Basketball and Mobile & Calendar sequences.

Figure 4 presents the variance of MPQM. It shows that the metric is capturing correctly the temporal perceptual quality variations characteristic of a CBR encoding as function of the bitrate.

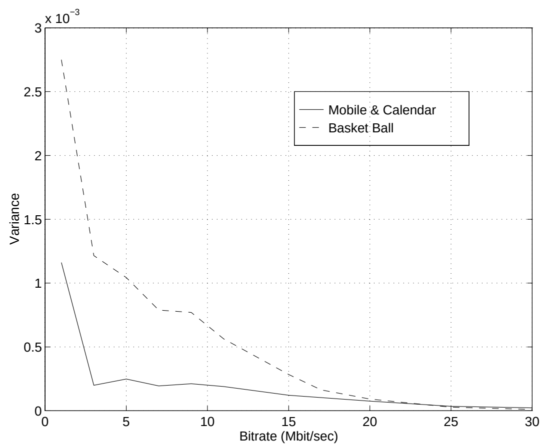

Figure 4: Variance of MPQM as a function of the bit rate for MPEG-2 encoding of the Basketball and Mobile & Calendar sequences.

6 Conclusion

In conclusion, we have applied three different video quality metrics, namely PSNR, ITS and MPQM to CBR MPEG-2 encoded standard test sequences and compared the results with subjective evaluations. We have shown that the MPQM is able to match very well the subjective assessment while the ITS metric and the PSNR are not.

References

[1] I. Dalgic and F.A. Tobagi. "Constant quality video encoding". *ICC-95*, June 1995.

[2] I. Dalgic and F. A. Tobagi. "A constant quality mpeg-1 video encoding scheme and its traffic characterization". *Picture Coding Symp.1996*, accepted for publication.

[3] W. K. Pratt. *Digital Image Processing.* Wiley and Sons NY (USA), 1974.

[4] T. W. Barnard. "A Literature Survey on Image Quality Evaluation". *The Perkin Elmer Corporation, Engineering Report ER-177*, 1971.

[5] W. C. Wilder. "Subjectively Relevant Error Criteria for Pictorial data Processing". *Purdue University, Engineering Report TR-EE 72-34*, Dec. 1972.

[6] CCIR recomm. 500-3. "Method for the Subjective Assessment of the Quality of Television Pictures".

[7] CCIR rep. 1082-1. "Studies toward the Unification of Picture Assessment Methodology".

[8] CCIR recomm. 813. "Methods for Objective Quality Assessment in relation to Impairments from Digital Codig of Television Signals".

[9] M. Ardito, M. Barbero, M. Stroppiana, and M. Visca. "Compression and Quality". *HDTV-94*, pp. B–8–2, Oct. 1994.

[10] M. Miyahara, K. Kotani, and V. R. Algazi. "Objective Picture Quality Scale (pqs) for Image Coding". *SID digest*, pp. 859–862, 1992.

[11] F. Lin and R. Mersereau. "A constant subjective quality MPEG encoder". *ICASSP*, pp. 2177–2180, 1995.

[12] I. Davies and D. Rose. "Automated Image Quality Assessment". *Human Vision, Visual Processing and Digital Display*, Vol. SPIE-1913, pp. 27–36, 1993.

[13] A. Webster, C. Jones, M. Pinson, S. Voran, and S. Wolf. "An Objective Video Quality Assessment System Based on Human Perception". *Human Vision, Visual Processing and Digital Display*, Vol. SPIE-1913, pp. 15–26, 1993.

[14] S. Wolf. "Features for Automated Quality Assessment of Digitally Transmitted Video". *U.S. dept. of commerce, Nat. Telecomm. and Inf. Adm. report*, pp. 90–264, June 1990.

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

[16] J. G. Daugman. "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 7, pp. 1169–1179, July 1988.

[17] Scott Daly. "The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity". In *Proceedings of SPIE*, Vol. 1616, pp. 2–15, 1992.

[18] Andrew B. Watson. "Perceptual-Component Architecture for Digital Video". *Journal of the Optical Society of America*, Vol. 7, No. 10, pp. 1943–1954, October 1990.

[19] 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.

[20] 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*, 1996. submitted paper.

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

[22] S. Western, K.L. Lagendijk, and J. Biemond. "Perceptual Image Quality based on a Multiple Channel HVS Model". *ICASSP*, pp. 2351–2354, 1995.

[23] J. L. Mannos and D. J. Sakrison. "The effects of a visual fidelity criterion on the encoding of images". *IEEE Transactions on Information Theory*, Vol. IT-20, No. 4, pp. 525–536, 1974.

[24] S. Comes and B. Macq. "Human Visual Quality Criterion". *SPIE Visual Communications and Image Processing*, Vol. 1360, pp. 2–7, 1990.

[25] Christian J. van den Branden Lambrecht and Olivier Verscheure. "Perceptual Quality Measure using a Spatio-Temporal M odel of the Human Visual System". In Proceedings of the IS&T Symposium on Electronic Imaging: Science and Technology, editors, *Digital Video Compression: Algorithms and Technologies 1996*, San Jose, CA, January 1996. The Society for Imaging Science and Technolog y. accepted for publication.

[26] Andrew B. Watson. *Handbook of Perception and Human Performance*, Chapter 6, Temporal Sensitivity. John Wiley, 1986.