A PERCEPTUAL PSNR BASED ON THE UTILIZATION OF A LINEAR MODEL OF HVS, MOTION VECTORS AND DFT-3D

M. Caramma, R. Lancini and M. Marconi
CEFRIEL-Politecnico di Milano
Via Fucini, 2. I-20133 Milano - ITALY
Tel: +39-02-23954-209; Fax: +39-02-23954-254
e-mail: rosa@cefriel.it

ABSTRACT

In this paper we propose two different methods to obtain a video sequence $PSNR$ taking into account of the subjective quality evaluation of the final users. These proposed "subjective" PSNRs are based on the spatio-temporal human sensitivity method suggested by Z. L. Budrikis [3] and on the experimental results given by Kelly [1] and Robson [2]. The first metric ($WPSNR_1$) consists in a PSNR weighed by the Motion Vectors typical of MPEG 1-2 [4] algorithms; the second one ($WPSNR_2$) is obtained by mean of the Discrete Fourier Transform applied to a three dimensional domain.
Experimental results show that these two WPSNRs are more related to the subjective quality perceived by the final user than the usual PSNR.

1 Introduction

In digital television, the need for storing and transmitting huge volumes of data makes the use of compression algorithms mandatory in the transmission environment and very cost-effective in the production environment. Nevertheless, in order to obtain good performance in terms of compression ratio, all the compression algorithms currently used can be considered lossy, that is to say they introduce different kinds of artefacts in the picture, with a consequent loss in picture quality. Up to now the picture quality of a new coding system has been evaluated by the votes of a panel of observers and, even if this method shows limits in term of time consumption and analysis of results, the subjective assessment is still now the only fundamental instruments to characterise the quality the new system will be able to offer. We may expect that in the future there will be a faster and faster development of new coding systems for image compression, and it would be an unrealistic hypothesis to carry out subjective assessment for every different coder available on the market. Objective measures are very simple to be implemented, easily repeatable in different laboratories and applicable to every kind of compression system (DCT based, wavelet based, fractal based etc.), the main drawbacks is that these measures do not correlate well with subjective quality measure. Even if on the average there is some correlation between objective and subjective picture quality, it is well known that two different sequences, subjected to the same compression system, for example MPEG-2 MP@ML, can show the same PSNR [3] and quite different subjective picture quality, and it is not unusual that the relationship between objective and subjective picture quality can be reversed. This behavior can be explained considering that averaging the measure on the frame means the loss of a large amount of information. In fact according to several studies concerning human perception, stimuli of the same amplitude/energy but different spatial frequencies are not perceived in the same way and, further, perception is different when a stimulus is included in flat spatial areas or in areas including edges, in a still portion of the image or in a portion containing movement, this loss of information usually nullify the effort to find a correlation between these values and subjective assessments.

Several laboratories are developing studies to find a correlation between objective measures of video impairments and human perception. Such new measures should replace subjective assessments; i.e. they should allow the evaluation of the performance of a given system without using any users. A possible solution to the problem consists in the development of a software tool intended to evaluate the performance of a coding algorithm in terms of perceived picture quality by using the coding error and some objective parameters related to the sequence. The definition of these objective parameters influencing the quality perceived by the observer is also fundamental in the refinement of coding techniques.

In this paper we propose two different methods to obtain the PSNR of a video sequence in order to take account of the subjective quality evaluation of the final users. The "subjective" PSNRs are based on the spatio-temporal human sensitivity method suggested by Z. L. Budrikis [3] and the experimental results given by Kelly [1] and Robson [2]. The first metric ($WPSNR_1$) consists in a PSNR based on the utilization of the Motion Vectors typical of MPEG 1-2 [4] algorithms; the second one ($WPSNR_2$) is obtained by mean of the

Discrete Fourier Transform applied to a three dimensional domain. Experimental results show that these two WPSNRs are more related to the subjective quality perceived by the final user than the customary PSNR. The paper is organized as follow: in the section II we show the model of Budrikis [3] based on the experimental results obtained by Kelly and Robson [1] [2]; in the sections III we describe how to use the model in the evaluation of $WPSNR_1$ using the Motion Vectors and WPSNR2 using the Discrete Fourier Transform into a three dimensional domain. Finally experimental results and conclusion are given in the section IV and V.

2 The Budrikis' model

Vision tests with sinusoidal stimuli go back at least to sixties, to Robson and Kelly [1]. The special interest in the sine-wave as test stimulus stems from the ease with which one can extrapolate from its results. In fact, provided a system is linear and time-invariant, Fourier analysis can be used to predict the system response to any input from its response to sinusoidal inputs. We are aware of the limitation of this approach because the visual system is neither linear nor time-invariant but nevertheless we will demonstrate that this coarse method is able to give some significant information about the perceived quality of the decoded video sequences with a very small computation cost. In [1] some experiments related with human sensitivity to spatio-temporal sinusoid are presented. The authors use a luminance pattern described by following formula:

$$L = L_0[1 + mcos(2\pi u_0 x)cos(2\pi f_0 t)]\,, \qquad (1)$$

which consists of a spatio-temporal sinusoid of amplitude m superimposed to a uniform background, L_0 is the average luminance, u_0 the spatial frequency, and f_0 the temporal frequency. A threshold value for m, beneath which the flickering pattern is no longer visible, is obtained for different frequencies. Its inverse can be interpreted as the human sensitivity to that particular frequency. The results are used in [3] to found a mathematical model for representing experimental results. A model, consisting of a difference between excitatory and inibitory terms, is proposed and separable functions of space and temporal frequencies are considered for these two terms. Finally authors consider direction-indipendent spatial terms and look for model's parameters minimizing mean-square deviations from experimental points presented in [1].

A mathematical expression of these concepts is then reported. First behaviour of Human Visual System (HVS) is assumed to be linear, so it can be described as follow:

$$C(x,y,t) = \int\int_A \int_0^\infty R(x-\xi, y-\eta, t-\tau)L(\xi,\eta,\tau)d\tau d\eta d\xi,$$

$$S(u,v,f) = \int\int_A \int_0^\infty R(x,y,t)e^{-2\pi j(ux+vy+ft)}dxdydt,$$

where $C(x,y,t)$ is the perceived distribution of luminance, $L(x,y,t)$ is the actual distribution of luminance in space-time, $R(x,y,t)$ is the impulsive-response of HVS and $S(u,v,f)$ the corresponding frequency-response. Then assumption on linear models leads to the following general structure for impulsive response:

$$R(\rho,t) = U_e(\rho)V_e(t) - U_i(\rho)V_i(t).$$

where e stays for excitatory and i for inihibitory. Finally one particular model taked into consideration in [3] is:

$$|S(\nu,f)| = A(e^{-\frac{f\tau_1}{2}}e^{-\nu\sigma_e} - ke^{-2\pi^2 f^2\tau_2^2}e^{-\nu\sigma_i}),$$

where $A, k, \tau_1, \tau_2, \sigma_e, \sigma_i$ are six undetermined parameters that author optimizes in [3]. The perspective view of this function is reported in Fig. (1).

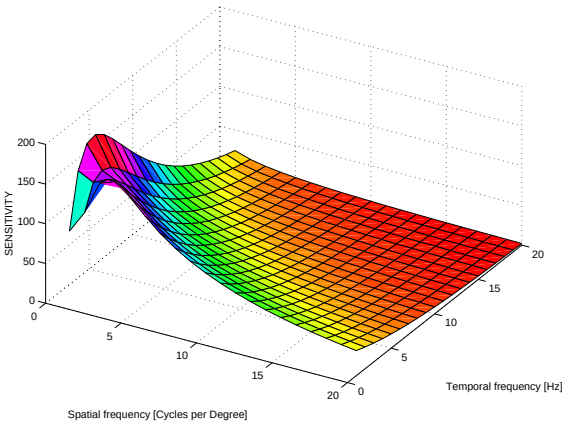

Figure 1: Amplitude value of Human Vision Frequency-Responce in Budrikis Modelization.

3 $WPSNR_1$ and $WPSNR_2$ evaluations based on the Budrikis' model

The most used measures are quantitative metrics based on a simple difference between frames, like MSE (Mean Square Error) and PSNR (Peak-Signal-to-Noise Ratio). They are easily analytically tractable, reliable and can be reproduced at any time. Their main drawback is that they do not correlate well with subjective quality evaluations. Even if on the average there is some correlation between objective and subjective picture quality, it is not unusual that different sequences, subjected to the same compression system, for example MPEG-2 MP@ML, show the same PSNR but quite different subjective picture quality. $PSNR$ is defined as follow:

$$PSNR(t) = 10Log_{10}\frac{255^2}{\epsilon^2(t)},$$

with

$$\epsilon^2(t) = \frac{1}{NM} \sum_{x,y}^{N,M} e^2(x,y,t),$$

where $e(x,y,t)$ is the coding error of the sequence, N and M are the dimensions of the frame. This quality measure is very tractable and used everywhere to estimate the coding performance.

The basic idea of this paper is to introduce informations given by HVS modelization exploited in [3] in evaluating a WPSNR (where W stays for weighted) defined by:

$$WPSNR(t) = 10Log_{10}\frac{255^2}{\epsilon_w^2(t)},$$

with

$$\epsilon_w^2(t) = \frac{1}{NM} \sum_{x,y}^{N,M} e_w^2(x,y,t),$$

and

$$e_w(x,y,t) = e(x,y,t) * R(x,y,t),$$

where $R(x,y,t)$ is the impulsive-responce of the HVS and e_w is the weighted coding error.

We can easily compute ϵ^2 in the frequency domain as

$$\epsilon_w^2 = \frac{1}{NMO} \sum_{x,y,t}^{N,M,O} e_w^2(x,y,t),$$

$$= \frac{1}{NMO} \sum_{x,y,t}^{N,M,O} (e(x,y,t) * R(x,y,t))^2,$$

$$= \frac{1}{NMO} \sum_{u,v,f}^{N,M,O} (E(u,v,f)S(u,v,f))^2.$$

where O represents the interval's time. Then two different approaches are followed. In the first one the coding error is considered frame by frame (i.e. $e(x,y,t_0)$), while in the second one a number of frames are grouped toghether in a three-dimensional macroblocks.

3.1 $WPSNR_1$ metric

Let's go in the details of the first approach, $WPSNR_1$. We divide each frame in macrobloks of 16x16 pixels as performed by MPEG algorithms, and we evaluate the Discrete Fourier Transform for each macroblock. To obtain temporal frequencies we use the motion vectors (MV) computed by a motion estimation algorithm in the formulas

$$f = \vec{f_s} \cdot \vec{v},$$
$$\vec{f_s} = (u,v),$$

where $\vec{(v)}$ is the velocity deduced by the MVs and $\vec{(f_s)}$ the spatial frequency. With the knowledge of the spatial and temporal frequencies, it is possible to find the value of sensitivity (see Fig. 1) and so the factor $R(x,y,t)$ in order to calculate $e_w(x,y,t)t)$.

3.2 $WPSNR_2$ metric

In the 2^nd approach, metric $WPSNR_2$ temporal frequencies are obtained whitout using motion vectors. A number of frame (16) is grouped toghether and three-dimensional macroblocks (i.e. *macrocubes*) are used. Finally Discrete Fourier Transorm in 3D is performed on each of them. Also in this case, the spatial and temporal frequency values allow us the calculation of $R(x,y,t)$.

4 Experimental results

In order to compare our technique with the usual PSNR we encoded different CCIR-601 sequences at a constant PSNR (30 dB). We evaluated $WPSNR_1$ and $WPSNR_2$ founding they are very similar to each other. The comparison between the usual $PSNR$ and the proposed metrics shown better performance of the last ones. In fact $WPSNR_1$ and $WPSNR_2$ are able to track better the behavior of the reconstructed subjective video quality.

A comparison between the PSNR and WPSNRs is shown in figures (2, 3, 4).

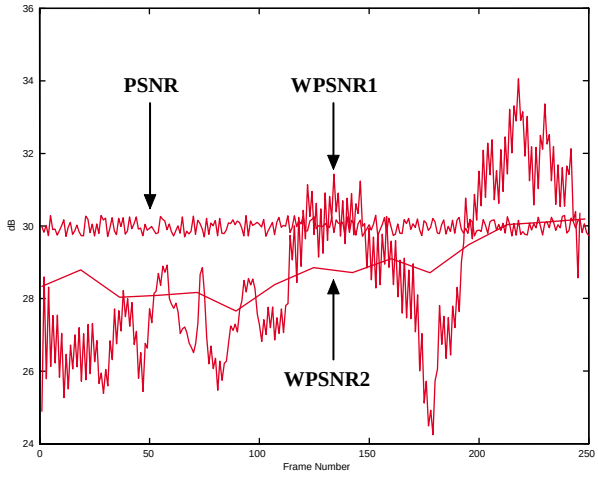

Figure 2: Comparison between $PSNR$ and $WPSNR_rs2$ of sequence *Dogs* coded at constant PSNR (30 dB).

Observing the decoded sequences it is possible to highlight a bigger similarity between $WPSNR_s$ and the actual perceived quality.

Sometimes $WPSNR_2$ seems to be better related with perceived quality than $WPSNR_1$. This is probably due to the incorrect estimation that affects the prediction of a number of motion vectors used in the computation of $WPSNR_1$.

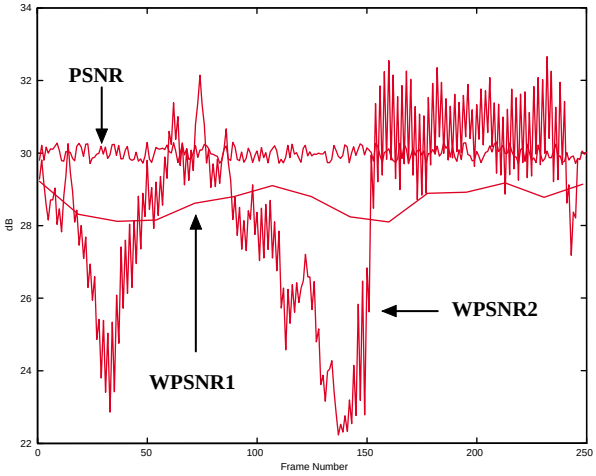

Figure 3: Comparison between $PSNR$ and $WPSNR_rs2$ of sequence $Rafting$ coded at constant PSNR (30 dB).

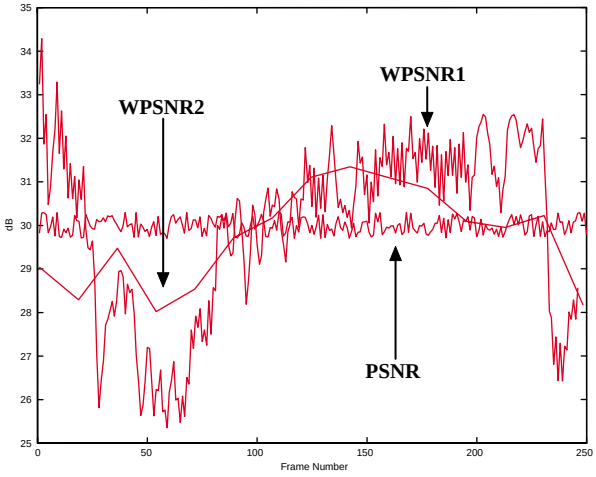

Figure 4: Comparison between $PSNR$ and $WPSNR_rs2$ of sequence Ski constant PSNR (30 dB).

5 Conclusion

In this paper we presented two approaches to introduce the model proposed in [3] in order to obtain a visual quality metric more related to the effective quality perceived by a human observer. The validity of our assumptions has been confirmed by experimental results on real video sequences.

Further research is being carried out to improve the accuracy of $WPSNR_1$ taking account of motion estimation reliability as presented in [5]. Finally a more accurate model of the HVS should be used taking account for its non-linear effects.

References

[1] Kelly, D.H. *J Stimulus pattern for Visual Research*, J. Opt. Soc. Am., 50, 1960, p.1115.

[2] Robson, J.J. *Spatial and Temporal Contrast Sensitivity Function of the Visual System*, J. Opt. Soc. Am., 56, 1966, p.1141-1142.

[3] Budrikis,Z.L. *Model Approximation to Visual Spatio-Temporal Sine-Wave Threshold Data*, The Bell System Technical Journal, November 1973.

[4] ISO/IEC JTC1/SC29/WG11 *Coding of moving picture and associated audio information*, ISO/IEC 13818-2, 1996.

[5] M. Caramma, R. Lancini, M. Marconi *Subjective Quality Evaluation of Video Sequences by using Motion Information*, Icip, 1999.