Objective Video Quality Metric based on Data Hiding

Mylène C.Q. Farias, *Member,* IEEE, Marco Carli, *Senior Member,* IEEE, and Sanjit K. Mitra, *Life Fellow,* IEEE

Abstract — *In this paper, a new no-reference (NR) objective metric based on data hiding is proposed. The metric has the advantage of being fast and not requiring knowledge of the original video contents. The proposed method uses a spread-spectrum embedding algorithm to embed a mark (binary image) into video frames. At the receiver, the mark is extracted and a measure of its degradation is used to estimate the quality of the video. We used data gathered from psychophysical experiments to help in the design of the video quality assessment system. We evaluated the visibility and annoyance of the impairments caused by the embedding algorithm and estimated the 'best' mark strength for a particular video. The performance of the proposed metric is estimated by measuring its ability to predict the Total Squared Error (TSE) of the host video and the Mean Observer Score (MOS) obtained from naive subjects in a psychophysical experiment. Experimental results show that the proposed metric had a good performance and a good correlation with the MOS*[1].

Index Terms — Video quality assessment, quality metric, artifacts, data hiding.

I. INTRODUCTION

The use of digital video has increased in recent years. Although there have been great advances in compression and transmission techniques, impairments are often introduced along the several stages of a communication system [1, 2]. The visibility and annoyance of these impairments are directly related to the quality of the received/processed video [3, 4]. For many applications, such as video conferences and broadcasting, it is important to have a good estimate of the quality of the material being received. Since in most applications humans are the ultimate receivers of the video material, the most accurate way to determine the quality of a video is to measure it directly using psychophysical experiments with human subjects [5]. Unfortunately, these experiments are too expensive and time-consuming to be a practical method for measuring video quality in real-time applications.

There is an ongoing effort to develop video quality metrics that are able to detect impairments and estimate their annoyance as perceived by human viewers. Most of the quality metrics proposed in the literature are Full Reference (FR) metrics. These metrics estimate the quality of a video by comparing the reference and impaired videos. Some examples include the works by Daly [6], Lubin [7], Watson [8], Wolf *et al.* [9], and Winkler [10]. A more complete survey of the available FR video quality metrics is presented in [11].

The major drawback of the FR approach is that a large amount of reference information has to be provided at the final comparison point. Also, a very precise spatial and temporal alignment of reference and impaired videos is needed to guarantee the accuracy of the metric.

Reduced Reference (RR) quality metrics are metrics that require only partial information about the reference video. In general, certain features or physical measures are extracted from the reference and transmitted to the receiver as side information to help evaluate the quality of the video. Metrics in this class may be less accurate than the FR metrics, but they are also less complex, and make real-time implementations more affordable. Some examples include the works of Webster *et al*. [12] and Brètillon *et al*. [13].

Requiring the reference video or even limited information about it becomes a serious impediment in many real-time applications. It is essential to develop no-reference (NR) video quality metrics that blindly estimate the quality of a video. It turns out that, although human observers can usually assess the quality of a video without using the reference, creating a metric that will implement this task is difficult and, most frequently, results in a loss of performance in comparison to the FR approach.

Most of the proposed NR metrics estimate annoyance by detecting and estimating the strength of commonly found artifact signals. For example, the metrics by Wu *et al.* and Wang *et al.* estimate quality based on blockiness measurements [14, 15], while the metric by Caviedes *et al.* takes into account measurements of 5 types of artifacts [16].

In this paper, we propose a new NR video quality metric that makes an unconventional use of data hiding system to blindly estimate the quality of a test video. In this approach, a binary mark is embedded into the original video frames before the compression and transmission stages. At the receiver, the mark is extracted and a measure of the degradation of the mark is used to estimate the quality of the test video. This type of

[1] This work is supported in part by a CAPES-Brazil fellowship, in part by a National Foundation grant No. CCR-0105404, and in part by an Italian National Research Council grant.

M.C.Q. Farias is currently with the Intel Corporation. (e-mail: mylene@ieee.org).

M. Carli is with Applied Electronics Dept., University of Roma 'TRE', Rome, Italy (e-mail: carli@uniroma3.it).

S.K. Mitra is with the Electrical and Computer Engineering Department, University of California, Santa Barbara, 93106 USA (e-mail: mitra@ece.ucsb.edu).

Contributed Paper
Manuscript received June 9, 2005

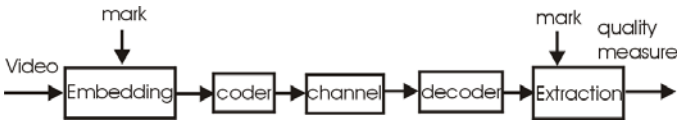

Fig. 1. Block diagram of the proposed video quality assessment system. A generic communication system is considered.

metric has the advantage of being fast and not requiring the use of the original video. A block diagram of the proposed quality assessment method is shown in Figure 1.

The metric proposed in this work follows the approach used by Campisi *et al.* where a spread spectrum embedding technique was used to assess the channel quality in wireless video transmissions [17]. In our work, we adopt a fragile data hiding system, i.e., the mark degrades with the host video. Based on previous results, we believe that this technique is adequate for the video quality assessment application [18]. We focus on estimating the quality of standard definition videos compressed using MPEG-2 and our target application is broadcasting. A detailed description of the embedding algorithm used in this work is presented in Section II.

The major contribution of this work consists of the use of data gathered from psychophysical experiments to help the design of the video quality assessment system and to test its performance. We first perform a psychophysical experiment to estimate the visibility threshold and mid-annoyance values of the impairments caused by the chosen embedding technique. In Section III, we present the details of this experiment and an analysis of the results obtained. Then, the video quality metric is designed using the knowledge acquired with the experiment and with simulation tests. In Section IV, we present the proposed video quality metric. The metric includes a step that estimates the 'best' mark strength based on the visibility threshold of the embedding impairments and the data hiding capacity of the host video [19]. The performance of the metric is evaluated by comparing the metric values with both the TSE of the test (host) videos and with the mean annoyance scores obtained (MOS) obtained from a second psychophysical experiment. Finally in Section V we draw our conclusions.

II. THE EMBEDDING ALGORITHM

Data embedding techniques have been used for several possible applications, such as, finger printing, multimedia indexing, and context base retrieval. A more recent application is the use of embedding techniques to estimate video quality at the receiver [18, 20, 21]. An embedding system for watermarking purposes has to satisfy three main constraints:

- Invisibility - the mark should not affect the perceptual quality of the video and should not produce noticeable distortions in the received data.
- Robustness - the mark cannot be altered by malicious (an attempt to alter the mark) or unintentional (compression, transmission or filtering) operations.
- Security - the mark may not be removed from the video, even if the embedding scheme is known.

However, the use of the embedding system with the purpose of estimating the quality of the host video changes the importance of these constraints. Invisibility, for example, is a very important constraint because our objective is drastically reduced if the mark is visible when the video is displayed on a computer or on a TV screen. Robustness, on the other hand, is not so important. In fact, if the mark is too robust, the extracted mark will not be affected unless the video is severely degraded. If the mark is too fragile, the extracted mark will be lost for small degradations making it difficult to differentiate between medium or highly degraded videos. Thus, for our application, the mark has to be semi-fragile and ideally it should degrade at around the same rate as the host video. Security is also not an issue for this application, since we are not trying to protect the video material.

Several embedding methods have been proposed in the literature. The mark can be inserted in the spatial domain [22] or in an *ad hoc* transform domain such as the DCT domain [23], the Fourier domain [24], or the wavelet domain [25]. In this work, we chose to insert the mark in the DCT domain, since this is the domain still used by many compression algorithms targeted at standard definition video TV formats.

Figure 2 depicts the block diagram of the embedding stage used by the proposed quality assessment method. The image mark **m**, a binary image, is embedded in each frame of the video using a spread-spectrum technique [23]. The embedding procedure can be summarized as follows. A pseudo random algorithm is first used to generate pseudo-noise (PN) images $\mathbf{p} = p(i,j,k)$ with values -1 or 1, and with a zero mean and Gaussian distribution. The indices i and j correspond to the horizontal and vertical positions, while k corresponds to the video index. A different pseudo-noise image is created for each frame of the video to avoid temporal summation.

The final mark to be embedded, **w**, is obtained by multiplying the binary mark image, **m**, by the PN image **p**:

$$w(i,j,k) = m(i,j) \cdot p(i,j,k). \qquad (1)$$

Notice that only one mark image, **m**, is used for all frames (k index), but the PN images vary from frame to frame. Then, the DCT transform, **LY**, of the logarithm of the luminance of the video frame, **y**, is computed:

$$\mathbf{LY} = \mathrm{DCT}(\mathbf{ly}) = \mathrm{DCT}(\log \mathbf{y}). \qquad (2)$$

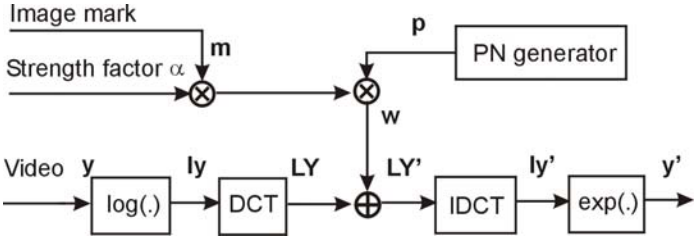

Fig. 2. Block diagram of the embedding stage of the video quality assessment system.

The logarithm was used for scaling purposes since this allows smaller values of scaling factor α to be used (see Eq. (3)) and, therefore, a smaller distortion. It has the disadvantage of causing a small increase in the computational complexity.

Then, the final mark **w** is added only to the mid-frequency DCT coefficients of the frame. The final mark, **w**, is multiplied by α before being added to the luminance DCT coefficients. After the embedding, the DCT coefficients are given by the following expression:

$$LY'(i,j,k) = \begin{cases} LY(i,j,k) + \alpha \cdot w(i,j,k), & 120 \le i,j \le 240, \\ LY(i,j,k), & \text{elsewhere.} \end{cases} \quad (3)$$

The range of frequencies where the mark is inserted is strongly dependent on the application. Inserting the mark in the low frequencies causes visible impairments in the image, while inserting it in the high frequencies makes it extremely fragile. For the purpose of assessing the quality of a video, the mid-frequencies are a good choice.

The scaling factor α is used to vary the strength of the mark. An increase of α increases the robustness of the mark, but also decreases the quality of the video. After the mark is inserted, the inverse DCT (IDCT) is calculated and then the exponential of the embedded coefficients video is taken. The video is then coded (compressed) and sent over the communication channel (see Figure 2).

The appropriate value of α depends on the type of application and video format. The design of an embedding system requires that an appropriate values of α be chosen for each video or set of frames. In Section III, we present a study of the influence of the values of α on the visibility and robustness of the mark. In Section IV, we present an algorithm for estimating the best α values for our application according to the visibility threshold and the content of the video.

Figure 3 shows the block diagram of the extraction stage of the video quality assessment system. If no errors are added by compression or transmission, the input of the extraction stage (Y'') is equal to the output of the embedding stage (Y'). On the other hand, if errors are added, then $Y'' = Y' + \eta$, where η represents the error signal. For the propose of explaining the extraction of the mark, we will assume that $Y'' = Y'$.

The process of extracting the mark from the received video is summarized as follows. The logarithm of the luminance of the received video, $\mathbf{y}''$, is first taken and its DCT is calculated. Then, we multiply the mid-frequency DCT coefficients where the mark was inserted by the correspon- ding pseudo-noise image, as given by the following equation:

$$\begin{aligned} LY''(i,j,k)\cdot p(i,j,k) &= LY(i,j,k)\cdot p(i,j,k) + \\ &+\alpha \cdot w(i,j,k)\cdot p(i,j,k) = \\ &= LY(i,j,k)\cdot p(i,j,k) + \\ &+\alpha \cdot m(i,j)\cdot p(i,j,k)\cdot p(i,j,k), \end{aligned} \quad (4)$$

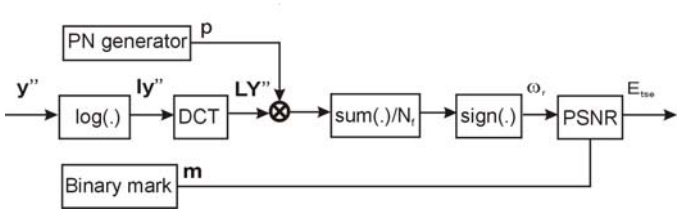

Fig. 3. Block diagram of extraction stage of the video quality assessment system.

for $120 \le i,j \le 240$. Since $p(i,j,k)\cdot p(i,j,k) = 1$ because $p(i,j,k)$ is either -1 or $+1$, the above equation becomes:

$$LY''(i,j,k)\cdot p(i,j,k) = LY(i,j,k)\cdot p(i,j,k) + \alpha \cdot m(i,j), \quad (5)$$

Synchronization is crucial at this step because the image mark can only be extracted if the same PN matrix used in the embedding is used in (4). Some bits of synchronization information can be easily embedded in the video to assure recovery.

The result of (5) is then averaged for a chosen number of frames N_f. This step is necessary to eliminate the noise (PN signal) introduced by the spread spectrum embedding algorithm. The extracted binary mark is obtained by taking the sign of this average, as given by the following expression:

$$\begin{aligned} m_r(i,j) &= \operatorname{sgn}\left(\frac{1}{N_f}\sum_{k=1}^{N_f} LY''(i,j,k)\cdot p(i,j,k) \right) \\ &= \operatorname{sgn}\left(\frac{1}{N_f}\sum_{k=1}^{N_f} LY(i,j,k)\cdot p(i,j,k) + \frac{1}{N_f}\sum_{k=1}^{N_f} \alpha \cdot m(i,j) \right) \\ &= \operatorname{sgn}\left(\frac{1}{N_f}\sum_{k=1}^{N_f} LY(i,j,k)\cdot p(i,j,k) + \alpha \cdot m(i,j,k) \right), \end{aligned} \quad (6)$$

Since the PN matrix has zero-mean, the sum $\Sigma_{k=1}^{N_f} LY(i,j,k)\cdot p(i,j,k)$ approaches zero for a large value of N_f. In general, for $N_f \le 10$ the mark is recovered perfectly, i.e., $\mathbf{m}_r = \mathbf{m}$. When errors are added by the compression or transmission, $Y'' = Y' + \eta$ and the extracted mark $\mathbf{m}_r$ is an approximation of **m**.

A measure of the degradation of the mark is given by the TSE of the extracted mark $\mathbf{m}_r$:

$$\mathrm{E_{tse}} = \sum_i \sum_j \left[m(i,j) - m_r(i,j) \right]^2. \quad (7)$$

The less the amount of errors caused by processing, compression or transmission, the smaller E_{tse} is. On the other hand, the more degraded the video, the higher E_{tse} is. Therefore, the measure given by E_{tse} can be used as an estimate of the degradation of the host video.

III. PSYCHOPHYSICAL EVALUATION OF EMBEDDING IMPAIRMENTS

The process of embedding data into a video may introduce undesired distortions or impairments that reduce the perceived quality of the video [26]. The main types of impairments introduced by embedding algorithms are:

- Flicker - Results from visible changes of the mark between consecutive frames.
- High Frequency Noise - Fundamental fingerprint of most embedding algorithms.

The visibility and annoyance of these impairments depend on several factors like the domain where the mark is being inserted, the embedding algorithm, and the strength of the mark (α). The stronger the mark, the more robust the detection is and, unfortunately, the more visible the impairments become.

To design a video quality system using data hiding techniques we have to make sure that (i) the embedding impairments are not visible and (ii) the mark is fragile and degrades at roughly the same rate as the the video. In this section we concentrate on (i), while in Section IV we will address (ii). In this section, we present the description of an experiment performed with the goal of estimating the detectability and annoyance of the impairments caused by the embedding algorithm.

A. Psychophysical Experiment Method

We used 20 test subjects drawn from a pool of students in the introductory psychology class at UCSB. The students were thought to be relatively naïve concerning video impairments and the associated terminology. They were asked to wear any vision correcting devices (glasses or contacts) that they normally wear to watch television. A Sony PVM-1343 monitor was used to display the test video sequences. The experiment was run with one subject at a time and lasted for approximately 50 minutes. Each subject is seated straight ahead of the monitor, located at or slightly below eye height for most subjects. The subjects are positioned at a distance of four screen heights (80 cm) from the video monitor.

The experimental session consists of five stages: instructions, training, practice trials, experimental trials, and interview. In the first stage, the subject is verbally given instructions. In the training stage, the subject is shown the original videos followed by examples of videos with the strongest impairments found in the experiment. These sequences represent the impairment extremes for the experiment and are used to establish the annoyance value range. The most annoying videos in the training stage should be assigned a value of '100'. In the practice stage, the test subjects run through several practice trials. The practice trials are identical to the experimental trials and are used to allow the responses to stabilize and to familiarize the test subject with the experiment [5].

The experimental trials stage is performed with the complete set of test sequences presented in a random order. In this stage, the subjects were asked to perform two different tasks: (1) detection and (2) annoyance. The detection task consisted of detecting impairments in the test sequences. The subjects are instructed to search each video for impairments. After each test sequence is played the subjects are asked "Did you see a defect or an impairment?." The subject is supposed to choose a 'yes' or 'no' answer. The annoyance task consisted of giving a numerical judgment of how annoying/bad the detected impairment is. The subject is instructed to enter a positive numerical value indicating how annoying the impairment is after each test sequence is played. Any defect as annoying as the worst impairments in the training stage should be given '100', half as annoying '50', ten percent as annoying '10', and so forth. Although the subjects are asked to enter annoyance values in the range of '0' to '100', they are told that values greater than '100' can be assigned if they think the impairment is worse than the most annoying impairments in the training stage.

After the experimental trials are complete, the test subjects are asked a few questions before they leave. These questions gather interesting information that cannot be gathered during the experiment. Nevertheless, they represent the subject's general impression of the set of test sequences and cannot be associated with specific sequences. However they are useful in guiding the design of future experiments.

B. Test Sequence Generation

To generate the test video sequences, we start with a set of five original video sequences of assumed high quality: 'Bus', 'Cheerleaders', 'Flower', 'Football', and 'Hockey'. This set of videos is commonly used in video quality research and is publicly available [27]. Representative frames of the original videos used in this work are shown in Figure 4. These videos are all five seconds long and are in ITU-R BT.601 format (formerly CCIR-601), i.e., the videos are 60 Hz (NTSC), 4:2:0 YCrCb format, 486 lines×720 columns. Two kinds of mark images used in our experiments were: the *Random* image and the *Logo* image shown in Figure 5. The marks are 88×88 size binary images.

Fig. 4. Sample frame of original videos 'Bus', 'Cheerleaders', 'Football', 'Flower', and 'Hockey'.

Fig. 5. Images used as marks: *Random* (left) and *Logo* (right).

Fig. 6. Zoomed version of the 10th frame of 'Cheerleaders': original (left) and embedded with mark *Logo* and α = 0.6 (right).

The test sequences are generated by embedding each original with both marks as described in the previous section. In order to be able to study the visibility and annoyance of embedding impairments, the contrast of these impairments (error patterns) must range from nearly imperceptible to highly annoying. This is obtained by varying the scaling factor (α) used in (3). The values of α used in this experiment were 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6. For illustration, in Figure 6 details of the 10th frame of the video 'Cheerleaders' with and without mark are shown. The picture on the left corresponds to the original video, while the picture on the right corresponds to the same video embedded with the mark *Logo*.

The total number of test sequences used in this experiment was 65, which includes 60 test sequences (5 originals × 6 strength factors× 2 mark images) plus the 5 original sequences.

C. Experimental results

Standard statistical methods are used to analyze the data provided by the test subjects [5]. The logarithm of the Total Squared Error (TSE) used in the analysis is given by:

$$\mathrm{TSE} = \sum_{k}\sum_{i}\sum_{j}\left[y(i,j,k) - y_o(i,j,k)\right]^2, \qquad (8)$$

where y is the impaired video and y_o is the original video, and i, j are the spatial coordinates, and k corresponds to the frame index.

To analyze the subjects' answers to detection tasks, we first convert the 'yes/no' answers to binary scores. The 'yes' is saved as '1', while 'no' is saved as '0'. The probability of detection (PD) of impairment is then estimated by counting the number of subjects who detect this impairment and dividing it by the total number of subjects. The mean observer score (MOS) is calculated by averaging the annoyance scores over all observers for each test sequence:

$$\mathrm{MOS} = \overline{Annoyance} = \frac{1}{L}\cdot\sum_{l=0}^{L} Annoyance(l), \qquad (9)$$

where *Annoyance* is the score reported by the *l*-th subject and L is the total number of subjects. We also calculated the sample standard deviation:

$$STD = \left(\frac{1}{L}\cdot\sum_{i=0}^{L}\left(Annoyance(i) - \overline{Annoyance}\right)^2\right)^{1/2}, \qquad (10)$$

and the internal standard error of $\overline{S}$:

$$\overline{STD} = STD/\sqrt{L} \qquad (11)$$

This analysis is valid under the assumption that the scores are independent. The confidence interval for the 'true' MOS of a test sequence is given by $\overline{S} \pm t_{L,\alpha/2}\overline{STD}$, where $t_{L,\alpha/2}$ corresponds to the *Student t* coefficient [28].

We divide the sequences into *test groups* composed of test sequences corresponding to the same original, with different levels of scaling factors and, consequently, annoyances. Using the probability of detection data, we can estimate the visibility detection threshold of the embedding impairments. The probability of detection as a function of the log(TSE) (*psychometric function*) is fitted using the Weibull function [5], which has an S-shape similar to the experimental data and is defined as:

$$PD(x) = 1 - 2^{-(S_T\cdot x)^{\kappa}}, \qquad (12)$$

where $PD(x)$ is the probability of detection, x is the $\log(TSE)$ of the sequence, S_T is the sensitivity, and κ is a constant that determines the steepness of the function. The 50% detection threshold in logarithmic error energy, x_T, is given by $1/S_T$.

Figures 7 and 8 depict the psychometric functions for the impairments caused by embedding the mark image *Logo* and *Random* in the videos 'Cheerleaders' and 'Football'. Each figure contains two curves, one for each of the mark images. As can be seen, for all figures both curves are very similar implying that the choice of a different mark image does not have a significant effect on the visibility of the impairments.

Columns 3 and 4 of Table 1 present the estimated 50% detection threshold for each test group in terms of TSE and $\log_{10}\mathrm{TSE}$. Overall, the threshold values do not change considerably over the test groups and remain practically constant when only the embedded image mark is changed. Because we only tested a finite number of α values, the exact α corresponding to the estimated visibility threshold (α_T) will most probably lie within one of the subintervals of the tested α values. In Column 5 of Table 1 the intervals containing the α_T or each test group are presented.

Table 1 also includes other curve fit parameters. The parameter sensitivity S_T (shown in Column 6) corresponds to the inverse of the log-threshold and therefore has an inverse behavior. The parameter κ (shown in Column 7), which represents the steepness of the probability of detection curve, varies between different test groups. This may be due to variations in the video content. The same impairment at the same strength will vary in visibility depending on the texture

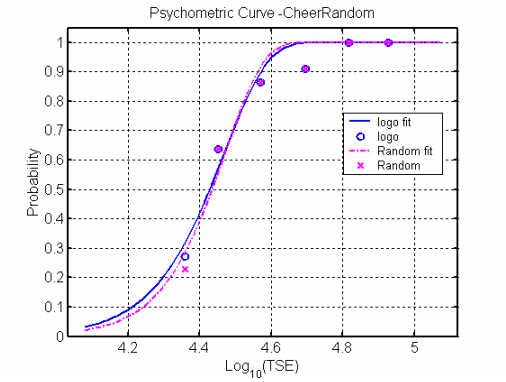

Fig. 7. Psychometric functions for embedding the marks *Logo* and *Random* into the video 'Cheerleaders'.

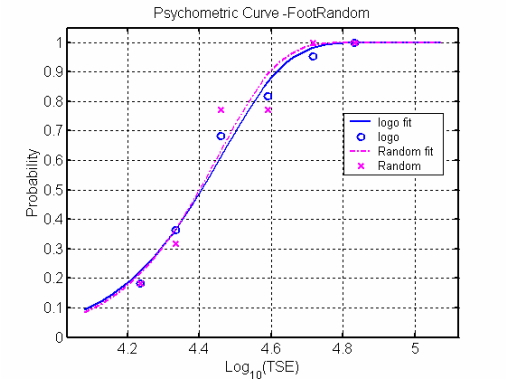

Fig. 8. Psychometric functions for embedding the marks *Logo* and *Random* into the video 'Football'.

and luminance characteristics of the background of the video. In particular, it was noticed that white and/or smooth backgrounds facilitate the visibility of the embedding impairments. For example, it was not possible to estimate the visibility threshold for the video 'Hockey' because more than half of all subjects saw the weakest impairment. It was also not possible to estimate the threshold for the video Flower embedded with the mark *Logo*.

With the annoyance data (MOS in (11)), we can estimate the mid-annoyance values of the embedding impairments. The MOS as a function of the $\log_{10}$ TSE (*annoyance function*), is fitted with the standard logistic function [5]:

$$y = y_{min} + \frac{(y_{max} - y_{min})}{1 + exp\left(-\frac{(x-\bar{x})}{\eta}\right)}, \qquad (13)$$

where y is the predicted annoyance and $x = \log_{10}$ TSE. The parameters y_{max} and y_{min} establish the limits of the annoyance value range. The parameter $\bar{x}$ (mid-annoyance value) translates the curve in the x-direction and the parameter η controls the steepness of the curve.

Figures 9 and 10 depict the annoyance functions for the impairments caused by embedding the mark image *Logo* and *Random*, respectively, into the videos 'Cheerleaders', and 'Football'. Each figure contains two curves, one for each mark. Again, the two curves look very similar, implying that the mid-annoyance values are not greatly affected by the choice of the image mark.

TABLE 1. PSYCHOMETRIC FUNCTION CURVE FIT PARAMETERS FOR EMBEDDING IMPAIRMENTS.

		Detection threshold (x_T)			Curve Fit Parameters	
Orig.	Mark	TSE	$\log_{10}$ TSE	α_T interval	S_T	κ
Bus	*Logo*	15488	4.19	$0.2 < \alpha_T < 0.3$	0.2407	27.71
Bus	*Random*	16218	4.21	$0.2 < \alpha_T < 0.3$	0.2378	31.17
Cheer	*Logo*	27542	4.44	$0.2 < \alpha_T < 0.3$	0.2259	37.66
Cheer	*Random*	28184	4.45	$0.2 < \alpha_T < 0.3$	0.2256	42.38
Flower	*Random*	39811	4.60	$0.1 < \alpha_T < 0.2$	0.2179	87.50
Foot	*Logo*	26303	4.42	$0.2 < \alpha_T < 0.3$	0.2270	25.68
Foot	*Random*	26303	4.42	$0.2 < \alpha_T < 0.3$	0.2273	27.38

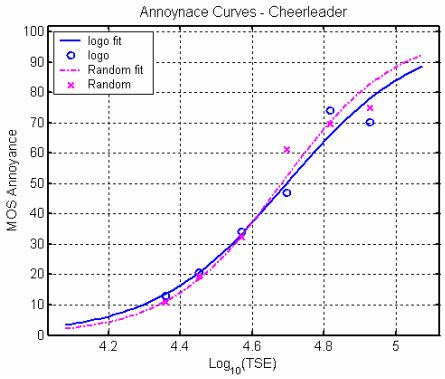

Fig. 9. Annoyance functions for embedding the marks *Logo* and *Random* into the video 'Cheerleaders'.

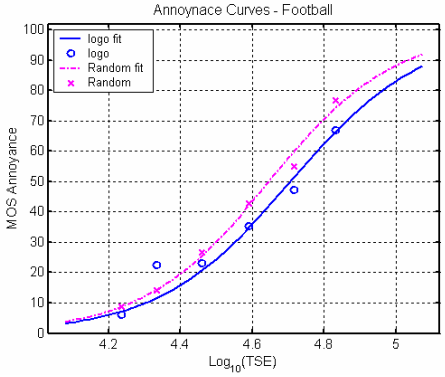

Fig. 10. Annoyance functions for embedding the marks *Logo* and *Random* into the video 'Football'.

Table 2 summarizes the fitting parameters for the annoyance functions. This table also shows the maximum, minimum, and the average MOS for each test group. As can be seen from Table 2, the steepness of the annoyance function, η (Column 7), does not vary significantly for different marks, but it does vary between test groups. The same is true for the parameter $\bar{x}$ (Column 6). The video 'Hockey' presented the highest mean, minimum, and maximum MOS values (Columns 3, 4, and 5, respectively). This can be explained by the fact that this scene contains large smooth white areas which greatly contrast with the embedding impairments (for an example of this effect, see Figure 6).

TABLE 2. ANNOYANCE FUNCTION FITTING PARAMETERS FOR EMBEDDING IMPAIRMENTS.

Test Sequence	Mark	Min	Max	Mean	$\bar{x}$	η	Res.
Bus	Logo	7.86	58.95	32.80	4.58	0.21	5.72
Bus	Random	9.80	63.05	32.71	4.55	0.2	2.91
Cheer	Logo	12.83	74.23	43.15	4.7	0.18	4.36
Cheer	Random	11.00	74.91	44.78	4.68	0.16	4.5
Flower	Logo	13.46	84.45	43.54	4.86	0.15	4.11
Flower	Random	13.25	85.82	43.92	4.85	0.14	4.5
Foot	Logo	6.25	66.91	33.55	4.71	0.18	5.06
Foot	Random	8.75	76.91	37.45	4.65	0.18	1.98
Hockey	Logo	30.00	95.55	65.19	4.61	0.12	2.11
Hockey	Random	41.00	96.95	68.64	4.58	0.12	6.46

IV. VIDEO QUALITY ASSESSMENT SYSTEM

The proposed system to blindly estimate the quality of a video is based on the assumption that the embedded image mark and the host video degrade at a similar rate. Therefore, the degradation of the extracted mark (E_{tse} in (9)) can be used as a measure of the quality of the distorted host video. We divided the design of the quality system in three steps. The first step consists in the development of an automated algorithm for estimating the 'best empirical' value of α for each video. This algorithm is an addition to the embedding stage presented in Section II and does not affect the extraction of the mark.

Once the adequate value of α is found, our second step consists of testing the proposed metric for different compression bit rates and comparing the results with a commonly used fidelity metric - the TSE. Comparing the values of the quality metric with TSE (or other type of fidelity metric) is common practice and it gives a good idea of the performance of the quality metric. Nevertheless, we believe that a more significant test consists of comparing the results of the quality metric with subjective results gathered from subjects. Therefore, the last step of the design of the video quality assessment system consists of comparing E_{tse} values with the MOS values obtained from a second psychophysical experiment that measured annoyance of MPEG-2 compressed videos [29].

According to the results presented in Section III, the choice of the mark is not critical to the visibility of the mark. Nevertheless, in order to obtain a better precision we decided to use a slightly larger mark image, which has less uniform areas and looks more like a natural image. We have chosen a (binary) dithered version of the image Lenna (seen in Figure 11) as our image mark for this section.

A. Automated System for Estimating α

While extracting the mark, it was noticed that some of the videos were more robust than others, i.e., for videos embedded with the same mark strength (α) and submitted to the same

Fig. 11. Dithered image, *Lenna*, used as an embedding mark for quality measurement.

degradations, the extracted marks resulted in different ranges of E_{tse}. This indicates that for different video contents different values of α are necessary. Therefore, an automated system for estimating appropriate values for α according to the video content has to be added to the embedding stage, as shown in Figure 12. No modifications are needed in the extraction stage, since this stage does not require the knowledge of the E_{tse} value.

For our application in video quality assessment system, good α values should correspond to E_{tse} values which:

- have a good range, i.e., the E_{tse} values are not concentrated in a very small range.
- are consistent, i.e., greater E_{tse} values should correspond to higher TSE or MOS, and smaller E_{tse} values should correspond to smaller TSE or MOS.

Our approach to find the best empirical values for α consisted of generating a set of test sequences embedded with different values of α and checking which α values satisfied the above conditions and were below the visibility threshold. To generate the test sequences we inserted the mark in each of the original videos with five different values of α (0.0125, 0.025, 0.050, 0.1, and 0.2). We compressed and decompressed each embedded video at compression rates ranging from 0.5 to 10 Mbps. Then, we extracted the mark and checked which values of α satisfied our criteria.

In Figure 13, we show the plot of E_{tse} versus host video TSE for different values of α and different compression levels corresponding to the original 'Football'. We can see from this graph that the value of α which better satisfied our criteria is 0.025. The E_{tse} values obtained for $\alpha = 0.025$ were not too low (mark being recovered too well) and the degradations

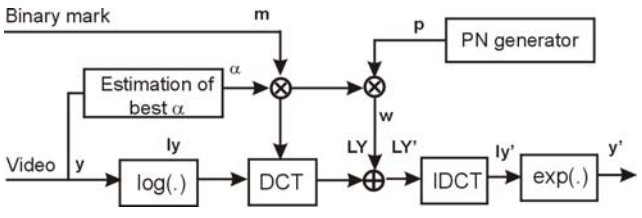

Fig. 12. Block diagram of the embedding stage of the video quality assessment system with an automatic estimation of best α.

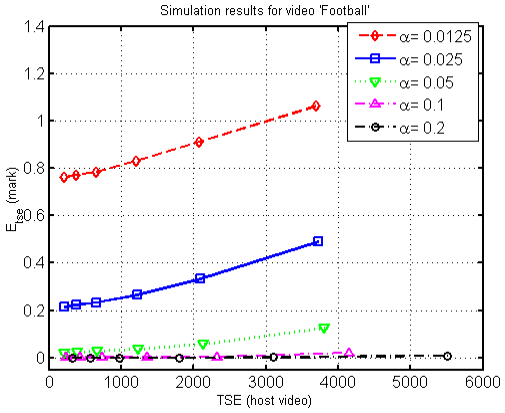

Fig. 13. E_{tse} **versus TSE of the host video for different values of α corresponding to the original 'Football'.**

were well captured by the curve. It also well below the visibility threshold. A similar procedure was used to find appropriate α values for the other originals.

In Column 5 of Table 3, we present the 'best emperical' α values according to the analysis over the test sequences and the results of the psychophysical experimental presented in Section III. As can be seen from Table 3, the α values corresponding to the visibility threshold (α_T in Column 4) and the 'best empirical' α values are not correlated.

The data hiding capacity of a video is given by the following expression [19]:

$$C = 0.5 \cdot \log\left(1 + \frac{\sigma^2_{mark}}{\sigma^2_{video}}\right) \qquad (14)$$

where σ^2_{mark} is the variance of the mark embedded and σ^2_{video} is the variance of the (host) video. The capacity and the standard deviation (σ) of the videos are shown in Columns 2 and 3 of Table 3, respectively. We can notice from Columns 3 and 5 that the standard deviation and the 'best empirical' α values are correlated.

In Figure 14 we plotted the 'best empirical' α versus the standard deviation for all videos. An exponential curve was fitted to this data:

$$\alpha_p(\sigma) = a \cdot \exp(b \cdot \sigma) \qquad (15)$$

where α_p is the predicted value for α, σ is the standard deviation of the host video. The values of a and b obtained from the fitting are 0.0162 and 0.1530, respectively. The correlation of the fit is 98.32%. Therefore, an automated system for estimating the value of α for each video can be implemented by simply measuring the standard deviation of the video frames and using (14).

B. Objective Metric Simulation Results for Compressed Videos

In this section, we present the simulation results of using the proposed metric to assess the quality of MPEG-2 compressed videos at several bit-rates. For this test we embedded the mark into five videos: 'Bus', 'Cheerleaders', 'Flower', 'Football',

TABLE 3. DATA HIDING CAPACITY, STANDARD DEVIATION, INTERVALS FOR THE VISIBILITY THRESHOLD, AND 'BEST EMPIRICAL' VALUES.

Test Sequence	Capacity	σ_{video}	α_T interval	'Best empirical' α
Flower	0.009	11.8955	$0.1 < \alpha_T < 0.2$	0.100
Bus	0.021	7.671	$0.2 < \alpha_T < 0.3$	0.050
Cheerleaders	0.025	7.0195	$0.2 < \alpha_T < 0.3$	0.050
Football	0.0718	1.2685	$0.2 < \alpha_T < 0.3$	0.025
Hockey	0.1549	0.8271	$0.0 < \alpha_T < 0.1$	0.0125

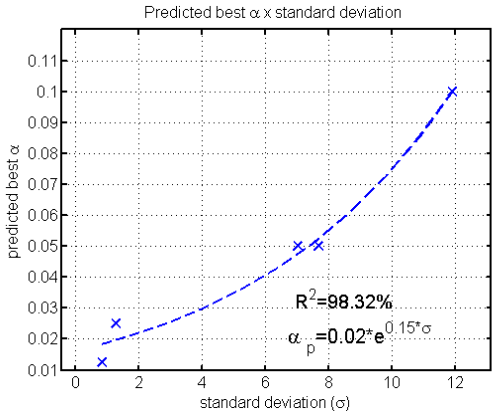

Fig. 14. Predicted E_{tse} **versus standard deviation of the host video.**

and 'Hockey'. The videos were embedded using the system shown in Figure 12, with α values calculated using (14). Once all the videos were embedded with the mark, they were compressed with an MPEG-2 codec at several bit-rates: 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 Mbps, respectively. The videos were then decompressed and the marks were extracted. Finally, the E_{tse} between the inserted and extracted marks were calculated.

Figures 15 and 16 depict the graphs of E_{tse} versus the $\log_{10}$ TSE between the original and degraded video. The graphs show that, as expected, the E_{tse} values increase monotonically with TSE of the host video. The range of the degraded video TSE varies depending on the video content, while the range of the proposed metric is always between 0

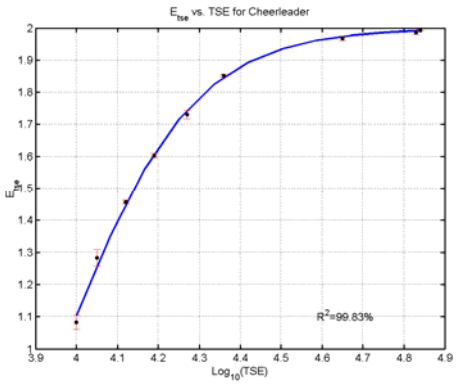

Fig. 15. E_{tse} **of the mark versus for video 'Cheerleaders'.**

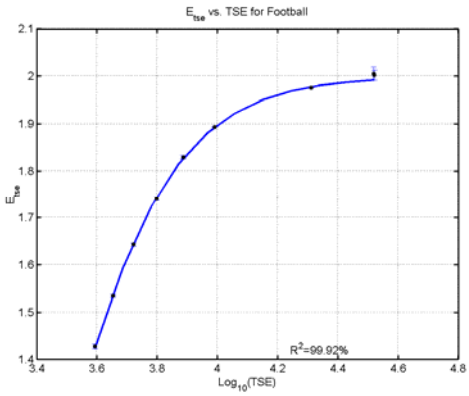

Fig. 16. E_{tse} **of the mark versus for video 'Football'.**

TABLE 4

FITTING PARAMETERS FOR CURVES OF E_{tse} **VERSUS** $\log_{10}$ TSE .

	log(*TSE*)	
Test Sequence	xmean	β
Bus	3.4631	0.2527
Cheerleaders	3.9671	0.1576
Flower	4.2532	0.2156
Football	3.4113	0.2018
Hockey	2.7802	0.1703

and 2. The curves were fit using a logistic function [5]. The graphs also show the fit correlation value (R^2) and the confidence intervals. The values of R^2 are all greater than 87% indicating a high correlation. The fitting parameters are shown in Table 4.

C. Objective Metric versus Mean Observer Score

The best performance metrics are full-reference metrics which make use of human visual system models, most of which are rather complex. Our metric is a no-reference metric with a much simpler approach. Our target application is on-line measurement of quality for broadcasting. We certainly do not expect to outperform full-reference metrics. Nevertheless, to be useful, our metric needs to demonstrate a significant correlation with the subjective data.

We compared the values obtained with our metric with the MOS values obtained from data gathered from a previous psychophysical experiment [29]. In this experiment, annoyance and detection of MPEG-2 impairments were measured. The same equipment, originals, and methodology described in Section II were used in this experiment. A total of 32 test subjects were used.

To create the test sequences we first generated sequences with high level of compression (1 Mbps). Then, we linearly combined the original video and the impaired video in different proportions. By varying their relative weights, we could weaken the artifact (allowing the original to dominate), strengthen the artifact (allowing the artifact to dominate), or even exaggerate the artifact (boost the difference between the artifact and original). This procedure guarantees that the appearance of the MPEG-2 impairments do not change, only their strength varies. The basic formula for combining the videos is:

$$Y = X_0 \cdot (1-r) + r \cdot X_1 \tag{16}$$

where Y is the result, X_0 is the original, X_1 is the degraded sequence, and r is the scaling factor. The defects are added only to pre-defined areas of the frames. All other areas are not distorted. The total number of test sequences used in this experiment was 95, which included 90 test sequences (5 originals times 6 strength factors times 3 defect zones) plus the 5 original sequences. The sequences are shown in a random order during the main experiment.

Figure 17 depicts the mean annoyance curves versus the E_{tse} for the set containing all videos. We can see that our metric is able to track the MOS of a video. Although the degradation will vary from video to video depending on their data hiding capacity, the E_{tse} provides consistent values and a good range for all the videos tested. E_{tse} also has a good correlation (0.8833) with the MOS values. The fitted line in the graphs is the quadratic function $y = -8.10x^2 + 52.35x$. The correlation coefficient for this fit is $R^2 = 80.11\%$.

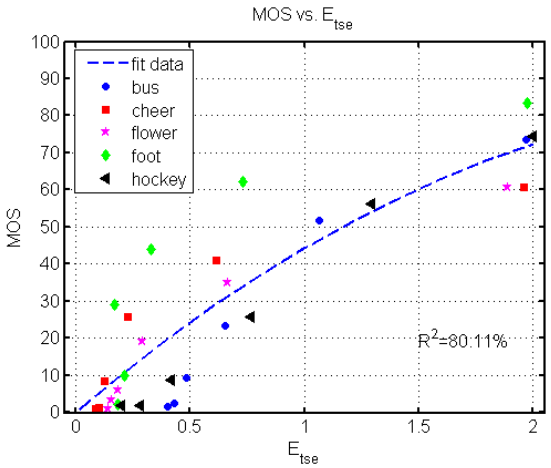

Fig. 17. MOS versus E_{tse} **for the complete set of test sequences.**

V. CONCLUSIONS

In this paper, a new NR objective metric using data hiding system is proposed. The metric has the advantage of being fast and not requiring the use of the original video. To evaluate the visibility and annoyance of the impairments caused by the chosen embedding algorithm, we performed a psychophysical experiment. The results of this experiment enabled us to study the relation between the visibility/annoyance of the mark and the mark strength. The system also includes an algorithm for estimating the 'best' strength of the mark based on its visibility and the data hiding capacity of the host video. The performance of the proposed metric is estimated by measuring its ability to predict the TSE of the host video and the MOS obtained from subjects in a psychophysical experiment. Although very simple, the proposed metric performed well and has a good correlation with the MOS.

REFERENCES

[1] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, pp. 247-78, October 1998.

[2] ITU Recommendation P.930, "Principles of a reference impairment system for video," 1996.

[3] Mylene C.Q. Farias, *No-Reference and Reduced Reference Video Quality Metrics: New Contributions*, Ph.D Dissertation, Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA, 2004.

[4] M.S. Moore, J.M. Foley, and S.K. Mitra, "Defect visibility and content importance: Effects on perceived impairment," *Image Communication*, vol. 19, pp. 185-203, February 2004.

[5] ITU Recommendation BT.500-8, "Methodology for the subjective assessment of the quality of television pictures," 1998.

[6] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital images and human vision*, Andrew B. Watson, Ed. Cambridge, Massachusetts: MIT Press, 1993, pp. 179-206.

[7] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital images and human vision*, Andrew B. Watson, Ed. Cambridge, Massachusetts: MIT Press, 1993, pp. 163-178.

[8] A. B. Watson, Hu James, and J. F. McGowan, "Digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, pp. 20-9.

[9] S. Wolf, M. H. Pinson, S. D. Voran, and A. A. Webster, "Objective quality assessment of digitally transmitted video," Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, pp. 477-82 vol, 9-10 May 1991.

[10] S. Winkler, *Vision models and quality metrics for image processing applications*, Ph.D. Dissertation, Signal Processing Laboratory, Ecole Polytechnique Federale de Lausanne, Lausanne, 2000.

[11] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, pp. 231-52, October 1999.

[12] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," Proc. SPIE Conference on Human Vision, Visual Processing, and Digital Display IV, San Jose, CA, USA, pp. 15-26, 1-4 February 1993.

[13] P. Bretillon, J. Baina, M. Jourlin, and G. Goudezeune, "Method for image quality monitoring on digital television networks," Proc. SPIE Conference on Multimedia Systems and Applications II, Boston, MA, USA, pp. 298-306, 20-22 Sept. 1999.

[14] H.R. Wu and M.; Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, pp. 317-320, Nov. 1997.

[15] Zhou Wang, A.C. Bovik, and B.L. Evan, "Blind measurement of blocking artifacts in images," Proc. IEEE International Conference on Image Processing, pp. 981-984, 10-13 September 2000.

[16] J. Caviedes and J. Jung, "No-Reference Metric for a Video Quality Control Loop," Proc. 5th World Multiconference on Systemics, cybernetics, and Informatics, pp. 290-5, July 2001.

[17] P. Campisi, M. Carli, G. Giunta, and A. Neri, "Blind quality assessment system for multimedia communication using tracing watermarking," *IEEE Transaction on Signal Processing, Special issue on Signal Processing for Data Hiding in Digital Media*, vol. 51, pp. 996-1002, April 2003.

[18] M.C.Q. Farias, S.K. Mitra, M. Carli, and A. Neri, "A comparison between an objective quality measure and the mean annoyance values of watermarked videos," Proc. IEEE International Conference on Image Processing, Rochester, NY, pp. 469 -472, 24-28 June 2002.

[19] M. Barni, F. Bartolini, A. De Rosa, and A. Abstract Piva, "Capacity of full frame DCT image watermarks," *IEEE Transactions of Image Processing*, vol. 9, pp. 1450-5, August 2000.

[20] O. Sugimoto, R. Kawada, M. Wada, and S. Matsumoto, "Objective measurement scheme for perceived picture quality degradation caused by MPEG encoding without any reference pictures," Proc. SPIE Conference on Human Vision and Electronic Imaging, San Jose, CA, USA, pp. 932-939, January 1998.

[21] M. Holliman and M. Young, "Watermarking for Automatic Quality Monitoring," Proc. SPIE Conference on Security and Watermarking of Multimedia Contents, San Jose, CA, USA, pp. 458-469, January 2002.

[22] N. Nikolaidis and I. Pitas, "Robust image watermarking in the spatial domain," *Signal Processing*, vol. 66, pp. 385-403, May 1998.

[23] M. Barni, F Bartolini, V. Cappellini, and A. Piva, "A DCT-domain system for robust image watermarking," *Signal Processing*, vol. 66, pp. 357-372, May 1998.

[24] I. Cox, J. Kilian, F. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. on Image Processing*, vol. 6, pp. 1673-1687, December 1997.

[25] H. Inoue, A.Miyazaki, and T. Katsura, "An image watermarking method based on the wavelet transform," Proc. IEEE International Conference on Image Processing, Kobe, Japan, pp. 296-300, October 1999.

[26] S. Winkler, E. D. Gelasca, and T. Ebrahimi, "Perceptual quality assessment for video watermarking," Proc. IEEE International Conference on Information Technology: Coding and Computing, Las Vegas, USA, pp. 90-4, April 2002.

[27] Video Quality Experts Group, "VQEG Subjective Test Plan," http://ftp.crc.ca/test/pub/crc/vqeg/.

[28] W. Hays, *Statistics for the social sciences*, 3 ed. Madison Avenue, New York, N.Y.: LLH Technology Publishing, 1981.

[29] M. S. Moore, J. M. Foley, and S. K. Mitra, "Comparison of the detectability and annoyance value of embedded MPEG-2 artifacts of different type, size, and duration," Proc. SPIE Conference on Human Vision and Electronic Imaging VI, San Jose, CA, pp. 90-101, January 2001.

Mylène C.Q. Farias (M'02) received her M.Sc. degree in electrical engineering in July 1998 from the Universidade Estadual de Campinas (UNICAMP)-Brazil, and her B.Sc. degree also in electrical engineering in January 1995 from Universidade Federal de Pernambuco (UFPE), Brazil amd her Ph.D. in electrical and computer engineering from the University of California Santa Barbara in 2004. She is currently working at the Intel Corporation, Chandler, AZ. Her current interests include video quality metrics, video processing, multimedia, watermarking, and information theory.

Marco Carli (SM'05) received the Laurea degree in Telecommunication Engineering from the University of Rome, "La Sapienza," Rome, Italy, in 1996. From 1997 he has been involved in European Union international programs in Distance Education. Since 2000, he has been a Visiting Researcher at the Image Processing Laboratory of the University of California, Santa Barbara, California for various periods. He currently holds an associate researcher position at the University of Rome, "Roma Tre" and is a lecturer for the courses on 'Multimedia Communications' and 'Telecommunication Systems' at the same university. His research interests are in the area of digital signal and image processing, in multimedia communications, and in security of telecommunication systems. He is a Senior Member of the IEEE, a member of the SPIE, and a reviewer for many conferences and for IEEE Transactions.

Sanjit K. Mitra (S'59-M'63-SM'69-F'74-LF-01) received the B.Sc. (Hons.) degree in physics from the Utkal University in 1953, the M.Sc. (Tech.) degree in radio physics and electronics from the Calcutta University in 1956, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1960 and 1962, respectively. He has been a Professor of Electrical and Computer Engineering at the University of California, Santa Barbara since 1977, where he served as the Chairman of the Department from July 1979 to June 1982. He has published over 600 papers in signal and image processing, twelve books, and holds five patents.

He has served IEEE in various capacities including service as the President of the IEEE Circuits and Systems Society in 1986 and as a Member-at-Large of the Board of Governors of the IEEE Signal Processing Society from 1996-99. He is a member of the U.S. National Academy of Engineeting, an Academician of the Academy of Finland, a member of the Norweigian Academy of Technological Sciences, a foreign member of the Croatian Academy of Sciences and Arts, and a foreign member of the Academy of Engineering of Mexico.