Quality assessment of H.264/AVC encoded video

Tomás Brandão*† , Luís Roque*‡ and Maria Paula Queluz*‡
*Instituto de Telecomunicações, Av. Rovisco Pais 1, Torre Norte, piso 10, 1049-001 Lisboa, Portugal.
Emails: {tomas.brandao, paula.queluz}@lx.it.pt
†ISCTE, Av. das Forças Armadas, 1649-026 Lisboa, Portugal.
‡Instituto Superior Técnico - Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisbon, Portugal.

Abstract— **In this paper, a new no-reference objective metric for the quality assessment of H.264/AVC encoded video is proposed and evaluated. Quality scores provided by the new metric are computed as a linear combination of simple features extracted from the video sequence received at the decoder. Using a set of 8 video sequences spanning a wide range of spatio-temporal activities, it is shown that the computed quality scores are well correlated with the ones resulting from subjective evaluation.**

I. Introduction

Quality assessment systems have a wide range of applications, from security services to entertainment, which includes digital television, internet video and in general the world of digital multimedia communications. It plays an important role in deciding the quality of service, network resources assignment and even to compare different service providers. However, the automatic evaluation of digital imaging systems quality is a complex task since it depends on a number of factors that contribute to what a viewer perceives as "video quality". Among these factors are the individual interests, quality expectations, viewing conditions and display type and properties [1].

In order to develop and standardize the required technology for assessing video quality, some organizations were formed. An example of that is the *Video Quality Experts Group* (VQEG), established in October of 1997. Video quality evaluation has thus become a relevant subject, which is also evidenced by the number of international conferences focused on this topic and products available (*e.g.*, video quality evaluation probes, known as *Witbe robots*, for measuring the quality of service offered by multimedia companies such as Portugal Telecom with MEO).

Evaluation of video quality can be achieved by subjective or objective metrics. The subjective video quality assessment is recognized as the most reliable mean of quantifying user perception since human beings are the ultimate receivers in most applications. The *Mean Opinion Score* (MOS), which is a subjective quality measurement obtained from a group of viewers, has been regarded for many years as the most consistent form of quality measurement. However, this quality measurement has some disadvantages – it is expensive for most applications, time consuming and cannot be executed automatically. Thus, in order to provide an automatic evaluation and monitoring of video data quality, objective metrics are required. By contrast to subjective measurements, the objective quality metrics are based purely on mathematical methods, from quite simplistic ones, like the *Peak Signal-to-Noise Ratio* (PSNR) and the *Mean Squared Error* (MSE), to sophisticated ones that exploit models of human visual perception and produce results far more consistent with the subjective evaluation.

According to the amount of the reference information required to assess the quality, objective video quality metrics are usually classified in three classes: *Full Reference* (FR), *Reduced Reference* (RR) and *No Reference* (NR). If the original video is totally available as well as the distorted video, the objective metrics are classified as FR. However, in many video service applications the reference video sequences are often not accessible; in that case, the metric is classified as NR if it is based only on the degraded video. In some cases, to improve the quality estimation, besides the distorted video some characteristics of the original video are also known and used, thus the objective metrics is categorized as RR metric. Comparatively to FR, few approaches were proposed for RR video quality assessment and even less for NR video quality evaluation. Some examples on this last group have been published in [2]–[4], for the purpose of PSNR estimation.

The work presented in this paper considers the two video quality assessment metrics mentioned previously, the subjective and the objective ones. The subjective tests have been conducted in order to obtain the MOS for a number of representative (in terms of spatial and temporal activities) video sequences, and after compressing those sequences with the H.264/AVC video coding standard [6]. These MOS values constitute the benchmark relatively to which the MOS predicted by the objective metric, and the metric itself, will be evaluated. Note that, among the different sources of video quality impairments, only those due to compression will be considered in this paper.

After the subjective tests having been carried out, a new NR objective video quality evaluation method is proposed and evaluated, the main purpose of which is to provide quality scores well correlated with the ones resulting from the subjective tests (MOS). Quality scores provided by the proposed metric are computed as a linear combination of simple features extracted from the video sequences at the decoder side.

This paper is organized has follows. After the Introduction, section II presents an overall description of the conditions and choices taken in order to perform the subjective tests sessions, as well as their results. Section III proposes a new NR objective video quality assessment method. Section IV depicts the results and a performance evaluation of the proposed metric, as well as the main conclusions of the paper.

II. Subjective quality evaluation

A. Methodology

The methodology followed in the subjective tests is standardized in ITU-R BT.500 [7] and ITU-T P.910 [8]. Recommendation ITU-R BT.500 has been, for long time, the reference for anyone who has to deal with subjective quality evaluation of television pictures, when displayed in the classical CRT screens. In this standard, several subjective evaluation methods are presented, covering different quality assessment scenarios. Recommendation ITU-T P.910 adapts Rec. ITU-R BT.500 to reduced picture formats (such as CIF, SIF and QCIF) and new types of display screens (*e.g.* LCD).

The subjective evaluation method followed in this work was the *Degradation Category Rating* (DCR) [8] also known in [7] as *Double Stimulus Impairment Scale* (DSIS). In this methodology, the observer is presented with video sequences organized in pairs: the first to be displayed is called the *reference* sequence (usually, the original) while the second is called the *test* or *impaired* sequence (for instance, the result of lossy encoding); it uses a five grade impairment scale, that reflects the observer´s judgment about the image impairment level: 1 – Very Annoying; 2 – Annoying; 3 – Slightly Annoying; 4 – Perceptible, but not Annoying; 5 – Imperceptible.

B. Assessment conditions

There are two essential elements for conducting the subjective quality evaluation sessions properly: the environmental viewing conditions and the test conditions. The main test conditions are [8]:

- Maximum test duration per session: 22 minutes
- Maximum number of observers per session: 2
- Viewing distance: 8 × the picture height shown in the screen

According to [8], at least 15 observers are needed in order to produce reliable and coherent results. In our tests, 22 observers (IST students) have been used. Before performing the subjective tests, they were screened for visual acuity and color blindness, using the Snellen Eye Chart and Ishihara's plates, respectively.

As for the environmental viewing conditions, three factors must be considered: the lighting, the ambiance noise and the quality and calibration of the display. The display and room characteristics used in the subjective tests, and listed below, are within the values recommended in [8].

- Height of the picture shown in the screen: 8 cm
- Viewing distance: 64 cm
- Background room illumination: 13.45 lux
- Peak luminance of the LCD screen: 95.8 lux
- Luminance of inactive screen: 2.23 lux
- Luminance of background behind the display: 10.15 lux
- Ratio of luminance of inactive screen to peak luminance: 0.023
- Ratio of luminance of background behind the display to peak of luminance: 0.14

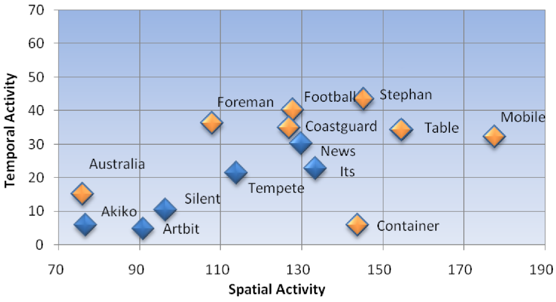


Fig. 1. Spatial and temporal activity measurements for a set of video sequences.

C. Selection of test material

In order to get meaningful and realistic tests results, it is important that a wide variety of video material is used during the tests. In particular, there are two relevant parameters which should be taken into account when choosing the test sequences: their spatial and temporal activities. Since, in order to avoid boring the observers, a small number of test sequences have to be used in the test sessions, it is important to choose a set of sequences that span a large range of possible values for these activities. The literature provides several different methods of measuring a video spatial and temporal activity. In this work, the methods recommended in [8] have been used:

- **Spatial activity**: the spatial activity measurement, in a very simplistic way, uses the two well known Sobel filters in order to compute the horizontal and vertical picture gradient. In order to obtain for each pixel a single measure, the gradient norm (the square root of the sum of the vertical and horizontal gradient squares) is obtained. The standard deviation of the gradient norm is then calculated for each frame, resulting in a time series of spatial activity of the sequence. In order to achieve a global value for the spatial activity, the maximum value in the time series is selected.
- **Temporal activity**: the temporal activity measure can be obtained computing the difference, pixel by pixel, between each two successive frames of the video sequence. After this procedure has been carried out, the standard deviation of the frames differences is computed. Similarly to what happens in the spatial activity, the global temporal activity value is computed as the maximum of these standard deviations.

However, because some sequences may present changes of camera perspective during video acquisition, or scene cuts, the resulting global activities could have a high value even if the sequence has a low temporal and/or spatial activity. In order to minimize and smooth this effect, before computing global values, the 95% percentile was applied to the temporal and spatial activities series. Results are presented in figure 1.

All video sequences are in CIF format (352×288 pixels), have 10 s duration, and all have a 30 Hz frame rate except

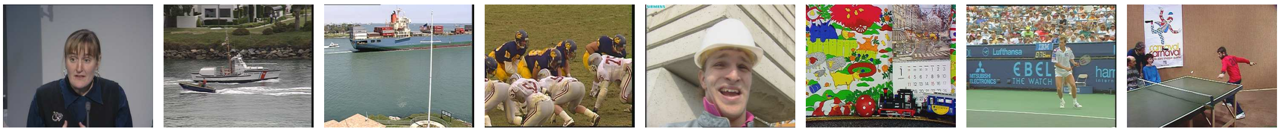


Fig. 2. Video sequences used in the subjective experiments. From left to right: *Australia*; *Coastguard*; *Container*; *Football*; *Foreman*; *Mobile & Calendar*; *Stephan*; *Table-tennis*.

Table I
Encoding bit rates for the sequences used in the tests.

Sequence	**Bit rates asked at the encoder**
Australia	32, 64, 128 and 256 kbit/s
Coastguard	64, 128, 256 and 512 kbit/s
Container	64, 128, 256 and 512 kbit/s
Football	256, 512, 1024 and 2048 kbit/s
Foreman	64, 128, 256 and 512 kbit/s
Mobile	64, 128, 256 and 512 kbit/s
Stephan	128, 256, 512 and 1024 kbit/s
Table	64, 128, 256 and 512 kbit/s

"Australia", which has a 25 Hz frame rate. From the full set of available video sequences, 8 sequences have been selected to be used in the subjective evaluation, which are presented in figure 2. This selection aims to cover a wide range of content with different spatio-temporal activity.

They were encoded using the H.264/AVC video coding standard using 4 different bit rates per sequence (summarized in table I). The result is a set of 32 encoded sequences, whose qualities have been judged by test participants. This set allows to test the HVS perception to different kinds of video qualities and to indirectly force the observers to use all available rating scale.

D. *MOS computation*

The mean opinion scores (MOS) are computed at the end of the session, based in the image quality assessment results given by all observers. In order to guarantee the coherence and the consistency of the results provided by the subjective tests, a statistical analysis (described in [7]-Annex 2) was applied to the assessment results. For each test condition, MOS values are computed by averaging the quality scores of the coherent observers, only.

III. Objective metric

A. *General description*

The objective metric proposed in this paper, represented in figure 3, results from combining a small set of simple features taken from the degraded video data subject to quality assessment. All the features in this set are known to influence video quality:

- *Video bit rate* – the bit rate of the encoded video. Generally, quality increases as bit rate increases, but not in a linear fashion (*i.e.*, the impact on quality of bit rate variations decreases as the bit rate increases). Thus, instead of considering the bit rate value directly as a model's input, its logarithm is used.
- *Mean square error estimate* – an estimation of the mean squared error between the reference and degraded video sequence is performed, using the algorithm proposed in [4]. This algorithm provides a no-reference PSNR estimate in a frame-by-frame basis, assuming that the video sequence is corrupted by quantization noise in the DCT domain (which is the case). Similarly to what has been done to the bit rate feature, the logarithm function has also been applied to the MSE estimated value.
- *Spatial and Temporal activities* – computed in the same way as described in section II-C, but using the received encoded sequence instead of the original one. Thus, it is assumed that the spatial and temporal activities of a video sequence are not significantly affected by the lossy encoding processed, which was confirmed experimentally.
- *Spatial activity and Temporal activity variances* - In order to account for activity changes along the video sequence, the variance of spatial and temporal activities, measured along time in a frame-by-frame basis, is also considered.

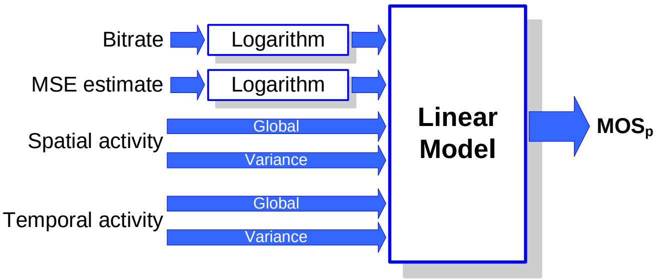


Fig. 3. Model used for MOS estimation.

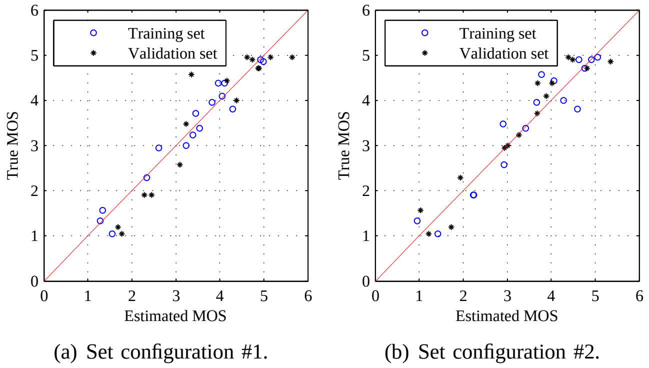
B. *Model parametrization*

The features described in the previous section are then combined using the linear model represented in figure 3, that computes a MOS estimate, MOS_p, according to:

$$MOS_p = \beta_0 + \sum_{k=1}^{N} \beta_k f_k, \qquad (1)$$

where f_k is the value of the k-th feature, β_k is the corresponding linear weight and N is the number of features. Using matrix notation, (1) can also be written as:

$$MOS_p = \mathbf{f}^{\mathbf{T}} \boldsymbol{\beta}, \qquad (2)$$



(a) Set configuration #1.

(b) Set configuration #2.

Fig. 4. MOS estimation results.

with

$$\mathbf{f}^{\mathbf{T}} = [1 \; f_1 \ldots f_N] \;\text{ and }\; \boldsymbol{\beta}^{\mathbf{T}} = [\beta_0 \; \beta_1 \ldots \beta_N] \,.$$

In order to find adequate values for vector $\boldsymbol{\beta}$, a training procedure is required. Thus, the set of encoded video sequences assessed during the subjective tests has been divided in two groups: training and evaluation sets.

One possible way to compute $\boldsymbol{\beta}$ is by minimizing the square error between MOS and MOS_p, for the video sequences in the training set. Assuming that the training set consists of K video sequences with their corresponding MOS values, K feature vectors will be extracted for training. The equation that gives $\boldsymbol{\beta}$ using the least square error criterion is given by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{k=1}^{K} (MOS^{(k)} - MOS_p^{(k)})^2 \right\}, \qquad (3)$$

which, in matrix form, can also be written as:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ [\mathbf{M} - \mathbf{F}\boldsymbol{\beta}]^{\mathbf{T}} [\mathbf{M} - \mathbf{F}\boldsymbol{\beta}] \right\}. \qquad (4)$$

with

$$\mathbf{F} = \begin{bmatrix} 1 & f_1^{(1)} & \ldots & f_N^{(1)} \\ 1 & f_1^{(2)} & \ldots & f_N^{(2)} \\ \vdots & \vdots & & \vdots \\ 1 & f_1^{(K)} & \ldots & f_N^{(K)} \end{bmatrix}, \text{ and } \mathbf{M} = \begin{bmatrix} MOS^{(1)} \\ MOS^{(2)} \\ \vdots \\ MOS^{(K)} \end{bmatrix}.$$

$\mathbf{F}$ is a $K \times N$ matrix, where each row contains the feature values taken from the k-th video sequence in the training set and $\mathbf{M}$ is a vector with the true MOS values. The least squares solution for $\boldsymbol{\beta}$ can be computed according to:

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}}\mathbf{M}. \qquad (5)$$

IV. Results and Conclusions

The 32 encoded video sequences evaluated in the subjective experiments have been randomly divided according to 16 sequences for training and another 16 for evaluation. In order to check for variability in the results, several combinations of training/validation set sequences have been used. Figure 4 depicts the results for two of these configurations. As can be observed from the figures, MOS estimates are close to their true values.

The performance of the proposed metric has been evaluated using the set of measurements proposed by the *Video Quality Experts Group* [9]. These measurements are usually address to as *prediction accuracy*, *monotonicity* and *consistency*. They are computed using the Pearson's correlation coefficient, Spearman's rank order coefficient and the outlier ratio, respectively. Additionally, the root mean square error (RMS) between MOS_p and MOS was also measured. The results are depicted in table II.

Table II
Evaluation of the proposed metric.

Performance measurement	**Config. #1**	**Config. #2**
Pearson correlation coefficient	0.953	0.952
Spearman rank order coefficient	0.960	0.958
Outlier ratio	0.125	0.125
Root mean square error	0.452	0.418

These results confirm the good performance of the algorithm proposed in this paper. When compared with the performance of [5], where a reduced-reference metric that also estimates quality scores through a linear combination of video features, the scheme proposed in this paper shows better results for all VQEG measurements (note, however, that the tests sequences were different).

References

[1] S. Winkler, "Video quality and beyond," in *proc. of EUSIPCO - European Signal Processing Conference*, Poznan, Poland, September 2007.

[2] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, "A method of estimating coding PSNR using quantized DCT coefficients," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 251–259, February 2006.

[3] A. Eden, "No-reference estimation of the coding PSNR for H.264-coded sequences," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 667–674, May 2007.

[4] T. Brandão and M. P. Queluz, "No-reference PSNR estimation algorithm for H.264 encoded video sequences," in *proc. of EUSIPCO - European Signal Processing Conference*, Lausanne, Switzerland, August 2008.

[5] T. Oelbaum and K. Diepold, "A reduced reference video quality metric for AVC / H.264," in *proc. of EUSIPCO - European Signal Processing Conference*, Poznan, Poland, September 2007, pp. 1265–1269.

[6] ITU-T, "Advanced video coding for generic audiovisual services," 2005.

[7] ITU-R, "Recommendation BT.500-11 – methodology for the subjective assessment of the quality of television pictures," 1974–2002.

[8] ITU-T, "Recommendation P.910 – subjective video quality assessment methods for multimedia applications," 1999.

[9] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," www.vqeg.org, Tech. Rep., August 2003.