Abstract — In this paper, two different strategies for the quality measurement of H.264/AVC and MPEG-2 encoded video are considered. These two video quality assessment approaches are commonly known as Subjective Methods and Objective Methods. With regard to the first one, a subjective video quality assessment test session was conducted, in order to achieve the Mean Opinion Score (MOS) for a group of representative video sequences. The subjective evaluation methods have been regarded, for many years, as the most reliable ones for measuring visual quality; however, they require a costly setup and appropriated viewing conditions. In order to provide an automatic evaluation and monitoring of video data quality, a MOS prediction model based on objective quality metrics is also proposed in this paper. Results show that the proposed metric is able to predict quality score well correlated with the subjective ones.

Keywords - Subjective Video Quality Metrics, Video Quality Assessment, Objective Video Quality Metrics, Mean Opinion Score.

I. INTRODUCTION

Quality assessment systems have a wide range of applications, from security services to entertainment, which include 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 Withe 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 methods. 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.

According to the amount of 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 practical video service applications the reference video sequences are 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 used, thus the objective metrics are categorized as RR metrics. Comparatively to FR, few approaches were proposed for RR video quality assessment and even less for NR quality evaluation.

The work presented in this paper has been organized taking into account 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 of a number of representative (in terms of spatial and temporal activities) video sequences, and after compressing those sequences with the MPEG-2 and the H.264/AVC video coding standards [6].

Besides the fact that subjective evaluation is the most recognized method for quantification of perceived quality, the attainment of the MOS for a number of representative compressed video sequences contributed to build a database of great interest for those working on the video quality evaluation field. The main reason of that significance is due to the fact that the majority of subjective results (e.g. those produced in MPEG groups) are only available for a restrict group of persons. Thus, the production of a database of video sequences and associated MOS, wins a new dimension of importance since the subjective results as well as of all type of information related with them, can be used in future works by people who has interest in video quality evaluation.

After the subjective tests having been carried out, a new NR objective 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 as follow. After the Introduction, an overall description of the conditions and choices taken in order to perform the subjective tests sessions, as well as their results, are described in section II. In section III, a new NR objective video quality assessment method is proposed. Section IV presents the results, including a performance evaluation of the proposed method. Finally, in section V, main conclusions are synthesized and further research topics are purposed.

II. SUBJECTIVE QUALITY EVALUATION

A. Methodology

The methodology followed in order to perform the subjective tests is standardized in the Recommendation ITU-R BT.500 [7] and in the Recommendation ITU-T P.910 [8] by the International Telecommunication Union group. 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 (“Subjective video quality assessment methods for multimedia applications”) adapts Rec. ITU-R BT.500 to multimedia applications such as videoconferencing, storage and retrieval applications and telemedicine applications, among others. The main difference between these two Recommendations is that Recommendation ITU-R BT.500 is focused on subjective assessment of video quality for television pictures, i.e., for large video formats; instead, Recommendation ITU-T P.910 is focused on subjective assessment of video quality for reduced picture formats (such as CIF, QCIF and SIF) and new types of display screens (e.g., LCD).

In this work, the evaluation method followed to perform the subjective tests was the Degradation Category Rating (DCR) [8] also known as Double Stimulus Impairment Scale (DSIS) in [7]. This methodology is appropriate for situations where the tests span the full range of impairments responsible for all visible degradation in the image. In this video quality evaluation method, the observer is presented with video sequences organized in pairs: the first to be displayed is called the reference sequence while the second is called the test or impaired sequence. The reference is the original, undistorted source sequence while the impaired sequence is a distorted version of the reference (for instance, the result of lossy encoding). After the sequences have been presented, the observer is asked to vote on the impaired sequence, but keeping in mind the first sequence as reference. The video quality assessment is performed using a five grade impairment scale that reflects the observer’s opinion about the image impairment level: 1 – Very Annoying; 2 – Annoying; 3 – Slightly Annoying; 4 – Perceptible, but not Annoying; 5 – Imperceptible.

B. Viewing and 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 × of the picture height shown in the screen (H)

In figure 2 the testing room used in this paper is schematically presented where the parameter H indicates the height of the video shown on the screen.

Other aspect taking into account in the subjective quality assessment tests is the observers’ selection. In order to produce reliable and coherent results, and in accordance with [8], at least 15 observers are needed, with increasing accuracy and consistency when this number increases. Our tests have been performed using 22 non-experts observers (IST students). This fact is relevant, since in general the public which consume the video material are commonly non-expert. Other reason that supports this choice is directly attached with the fact that the non-experts are not concerned with television picture as part of their normal work. Therefore, the non-experts do not have a pre-determined way of watching a video sequence as the experts have. In [7], preliminary findings suggest that non-experts observers may yield more critical results with exposure to higher quality transmission and displays technologies.
In the subjective quality assessment, the human eye has a special importance. So, before performing the subjective tests, the observers were submitted to ophthalmologic tests in which they were screened for visual acuity and color blindness, using the Snellen Eye Chart and Ishihara’s plates, respectively (Figure 3).

![Figure 3. (a) Ishihara Test plate; (b) Snellen Eye Chart](image)

Another aspect that should be taking into consideration when planning a subjective tests session is that different environments with different viewing conditions can affect the experimental results. In special, there are three factors that must be considered when performing the subjective tests: the lighting, the ambience noise and the quality and calibration of the display. The display room characteristics used in the subjective tests, listed in table 1, are in accordance with the values recommended in [8].

<table>
<thead>
<tr>
<th>Table 1. Display and Room’s conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameters</strong></td>
</tr>
<tr>
<td>Height of the picture shown in the screen (H)</td>
</tr>
<tr>
<td>Viewing distance</td>
</tr>
<tr>
<td>Background room illumination</td>
</tr>
<tr>
<td>Peak luminance of the screen</td>
</tr>
<tr>
<td>Luminance of inactive screen</td>
</tr>
<tr>
<td>Luminance of background behind the display</td>
</tr>
<tr>
<td>Ratio of luminance of inactive screen to peak luminance</td>
</tr>
<tr>
<td>Ratio of luminance of background behind the display</td>
</tr>
</tbody>
</table>

Regarding the sessions structure, it should not exceed half an hour [7], since if this does not happen the observer gets tired and, as consequence, the results will not be coherent. The evaluation sessions are divided in two parts: warm-up session and the real test session. According to [7] the warm-up session is presented to the observers initially, before the real test session begins, as it is possible to see in figure 4.

Figure 4. Test session structure

The intention of the warm-up phase is to guarantee the observer’s opinion stabilization and to define in his/her mind some video quality boundaries. It is important to add that the data issued from this phase should not be taken into consideration for further analysis. The results of the second phase – actual evaluation – are the truthful goal of all entire subjective quality evaluation sessions. It is from this second stage, that the observers’ results will be taken into account in order to calculate the Mean Opinion Score (MOS).

C. Test material selection

The subjective quality assessment results strongly depend on the videos scene or sequence content selected to be viewed by the observers. In consequence, the selection of test material must be done carefully. 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 scenes: their spatial and temporal activities. In fact, according to [9], the human visual perception of video content is determined by the video spatial information, as well as by the type, direction as well as the speed of movement, or temporal activity. 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.

The kernel of the Sobel filters is presented in figure 5.
Figure 5. Sobel filters. (a) Sobel filter responsible for detecting horizontal pixel differences; (b) Sobel filter responsible for detecting vertical pixel differences.

Figure 6 shows the resulting gradient norm for a video sequence frame.

(a) (b)

Figure 6. (a) Original video frame; (b) Corresponding gradient norm image.

- **Temporal activity**: The temporal activity measurement can be obtained computing the difference, pixel by pixel, between each two successive frames of the video sequence. This process is repeated for all video sequence frames. After this procedure has been carried out, the standard deviation of the frames differences is computed. Similarly to what it happens in the spatial activity, the global temporal activity value is computed as the maximum of these standard deviations.

Figure 7 presents, in the right side, two consecutive frames of the original video and, in the left side, the resulting difference between the two original frames.

Figure 7. Temporal activity measurement process in a video sequence.

Figure 8. Spatial-temporal activity measurements for a video sequence set.

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. Figure 8 presents the computed activities for a set of video sequences commonly used by the video coding community. All sequences are in CIF format (352×288 pixels), have 10 s duration and, except Australia, which has a frame rate of 25 Hz, all have a 30 Hz frame rate.

In accordance with figure 8 and taking into account that the video sequences must span a large portion of the spatial-temporal information, eight sequences have been selected to be used in the subjective evaluation. The selected sequences are presented in figure 9.

Figure 9. Video sequences used in the subjective tests: (1) Australia; (2) Coastguard; (3) Container; (4) Football; (5) Foreman; (6) Mobile; (7) Stefan; (8) Table-tennis.

The selected sequences were compressed using two broadly used compression standards, the H.264/AVC and the MPEG-2. In order to display to the observers different types of video quality, each video sequence was compressed with four different bitrates (summarized in table 2). The set of resulting encoded sequences will allow to test the HVS perception to different kinds of video qualities and to indirectly force the observers to use all available rating scale.
Table 2. Encoding bit rates for the sequences used in the tests for: (a) H.264 and (b) MPEG-2.

(a)  
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Trial 1 [kbit/s]</th>
<th>Trial 2 [kbit/s]</th>
<th>Trial 3 [kbit/s]</th>
<th>Trial 4 [kbit/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>Coastguard</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>Container</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>Football</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>2048</td>
</tr>
<tr>
<td>Foreman</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>Mobile</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>Stephan</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>Table</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
</tbody>
</table>

(b)  
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Trial 1 [kbit/s]</th>
<th>Trial 2 [kbit/s]</th>
<th>Trial 3 [kbit/s]</th>
<th>Trial 4 [kbit/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>Coastguard</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>2048</td>
</tr>
<tr>
<td>Container</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>Football</td>
<td>512</td>
<td>1024</td>
<td>2048</td>
<td>4096</td>
</tr>
<tr>
<td>Foreman</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>2048</td>
</tr>
<tr>
<td>Mobile</td>
<td>256</td>
<td>512</td>
<td>1024</td>
<td>4096</td>
</tr>
<tr>
<td>Stephan</td>
<td>512</td>
<td>1024</td>
<td>2048</td>
<td>4096</td>
</tr>
<tr>
<td>Table</td>
<td>256</td>
<td>512</td>
<td>2048</td>
<td>1024</td>
</tr>
</tbody>
</table>

D. MOS computation

Based on the image quality assessment results given by all observers in each one of the sessions, the mean opinion scores (MOS) were computed at the end of the subjective quality assessment sessions. 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 performed to the assessment results. After that, for each test condition, new MOS values were computed by averaging the quality scores of the consistent observers, only.

The video material used to perform the subjective tests, such as the original video sequences, the H.264 and MPEG-2 compressed videos, as well as the tests results – the opinion scores and the MOS – are available for those who are interested on the video quality evaluation field at http://amalia.img.lx.it.pt/~tgsb/H264_test/.

III. OBJECTIVE QUALITY METRIC

A. General overview

The subjective tests are particularly important in video quality evaluation since they provide the means to quantify quality as it is perceived by the viewers. However, they are not suitable for monitoring video data quality. Recently, the increasing success of digital video has motivated the research of objective quality evaluation metrics. These metrics aim to assess the quality of a broadcasted video as it is perceived at the user-end, automatically and in a real time basis. In order to validate the performance of an objective quality metric, the human assessment must also be taken into account. In this section an automatic objective video quality assessment is proposed. The developed objective metric combines a set of features taken from the degraded video data. From the analysis of the subjective results it was possible to relate some video features with the correspondent video quality grades given by the observers. Some features which are known to have influence on video quality are:

- **Video bit rate** – As expected, MOS increases with the bit rate of the encoded video sequences, as it is possible to notice in figure 10.

![Figure 10. MOS evolution with the video bit rate of some video sequences](image)

Observing the figure, it can be seen that the MOS evolution with the bit rate is not linear: for higher bitrates a large variation on the bit rate does not lead to a significant variation on the MOS; on the other hand, for lower bitrates a small bit rate variation can conduct to a large MOS variation. Thus, instead of considering the bit rate value directly as a model’s input, its logarithm is used.

- **Mean square error estimate (MSE)** – Similarly to the video bit rate, the MSE also presents a close relationship with the correspondent video quality grades given by the observers and consequently MOS, as shown in figure 11.

![Figure 11. MOS evolution with the MSE of some video sequences](image)

In this case, MOS values are roughly inversely proportional to the MSE values. Thus, the higher is the difference (MSE) between each frame of the original and the encoded video, the lower will be the grade given by the observers (MOS).
Since in many practical video service applications the original video sequences are often not accessible, an estimation of the mean squared error between the reference and degraded video sequence is performed, using the PSNR estimation model proposed by Brandão et al. [3]. 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).

- **Mean square error estimate variance** – In order to give a better and precise video sequence description along time, the MSE variance is also considered.

- **Spatial and Temporal activities** – The spatial-temporal activity of a video sequence has also an important role in human video quality evaluation, *i.e.*, the HVS is largely influenced by the movement and the edges present in a video. Figure 12, presents the resulting MOS values evolution with this spatial and the temporal activities for a set of video sequences encoded at two different bitrates.

  By observing figures 12.a) and b), it is possible to state that, at the same bit rate, MOS values of video sequences with a large spatial-temporal activity are negatively affected when compared to video sequences with reduced spatial-temporal activity. In fact, the spatial-temporal activity of a video sequence has an important role in its video quality and, indirectly, in its video quality evaluation provided by the observers. The HVS is largely influenced by the movement and texture contents in the video. Thus, when a video sequence characterized by a large spatial-temporal activity is encoded at low bitrates, its quality is more affected than the videos which have reduced spatial-temporal activity.

- **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. This is especially important if the video sequence presents high variations of these features, since the global value fails to explain completely how these features evolve during the sequence.

Taking into consideration this set of features, two approaches for MOS prediction model are proposed. The objective is to develop MOS prediction models that are based only in Non-Reference (NR) features, *i.e.*, computed at the receiver side from the received video sequence. The features considered for the first approach are:

- Bit Rate (BR);
- Global Spatial Activity (gSA);
- Global Temporal Activity (gTA);
- Spatial Activity Variance (vSA);

This first MOS prediction model can be formally described by

\[ \hat{MOS} = f_{1}(BR, gSA, vSA, vTA, vSA) \]  \hspace{1cm} (1)

The second approach extends the first one by also considering the MSE feature. It is important to remark that this second approach has a higher computational complexity than the first one, resulting from the inclusion of an algorithm that estimates the MSE [3]. Thus, this more complex MOS prediction model will be based on the following features:

- Bit Rate (BR);
- Global Spatial Activity (gSA);
- Global Temporal Activity (gTA);
- Spatial Activity Variance (vSA);
- Temporal Activity Variance (vTA);
- Global MSE (gMSE);
- MSE Variance (vMSE);

Figure 12. MOS relationship with spatial and temporal activities for video sequences (table 2) with two different bitrates: (a) 128 kbit/s and (b) 1024 kbit/s.
Thus, the second approach can be formally described by
\[
\text{MÖS} = f_2(BR, gMSE, vMSE, gTA, gSA, vTA, vSA)
\]  \hspace{1cm} (2)

Although the inclusion of the MSE as a feature increases the system and computational complexity, it is of interest to evaluate its influence in the accuracy of the MOS estimation. Figure 13 depicts the MOS estimation model based on the two approaches described in (1) and (2), the low complexity and the high complexity based models, respectively.

![Figure 13. General MOS estimation model overview: (a) low complexity based model; (b) high complexity based model.](image)

**B. Regression model**

The MOS prediction model developed in this paper explores the influence that a set of independent variables (features) described in the previous section have on the MOS, and uses a model capable of combining all the pertinent information available by using a linear regression equation.

The proposed linear model that computes a MOS estimate, MÖS, based on a set of simple features, is given by,
\[
\text{MÖS} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i,
\]  \hspace{1cm} (3)

where \(x_i\) is the value of the \(i\)-th feature, \(\beta_i\) is the corresponding linear weight and \(N\) is the number of features. Using matrix notation, (3) can also be written as:
\[
\text{MÖS} = x^T \beta,
\]  \hspace{1cm} (4)

where \(x^T = [x_1 \ldots x_N]\) and \(\beta^T = [\beta_0 \beta_1 \ldots \beta_N]\).

Mathematically, the two linear regression models described in section III-A, the low and the high complexity models, are respectively given by
\[
\begin{align*}
\text{MÖS}_\text{gMSE} &= \beta_0 + \beta_1 \times \log(BR) + \beta_2 \times gSA + \beta_3 \times gTA + \beta_4 \times vSA + \beta_5 \times vTA \\
&+ \beta_6 \times vTA + \beta_7 \times \text{MÖS}_{\text{MSE}}
\end{align*}
\]  \hspace{1cm} (5)

where \(\text{MÖS}_{\text{gMSE}}\) and \(\text{MÖS}_{\text{MSE}}\) are the MOS estimation taking into account only one feature: the global MSE and the MSE variance, respectively. The \(\text{MÖS}_{\text{gMSE}}\) and \(\text{MÖS}_{\text{MSE}}\) are given, respectively, by:
\[
\begin{align*}
\text{MÖS}_{\text{gMSE}} &= \beta_{0gMSE} + (\beta_{1gMSE} \times gMSE) + (\beta_{2gMSE} \times gMSE)^2 \\
\text{MÖS}_{\text{MSE}} &= \beta_{0vMSE} + (\beta_{1vMSE} \times vMSE) + (\beta_{2vMSE} \times vMSE)^2
\end{align*}
\]  \hspace{1cm} (6)

In order to develop a model capable of predicting the MOS, it is important to determine the adequate values for vector \(\beta\). In this step, the importance of the subjective tests is once again referred since the set of encoded video sequences assessed during the subjective tests has been divided in two groups: a training set used to calibrate the regression weights, \(\beta\) and an evaluation set used to estimate the MOS and consequently testing the model accuracy.

Regarding the regression weights computation, one possible method to compute \(\beta\) is by minimizing the square error between MOS and MÖS, for the training set video sequences. Since the training set is based on \(K\) video sequences with their corresponding MOS values, \(K\) feature vectors will be extracted for training. Thus, using the least square error criterion, the vector \(\hat{\beta}\) is given by:
\[
\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{j=1}^{K} (\text{MÖS}^{(j)} - \text{MÖS}^{\text{MÖS}}^{(j)})^2 \right\}
\]
\[
\begin{align*}
&= \arg\min_{\beta} \left\{ \sum_{j=1}^{K} \left( \beta_0 + \sum_{i=1}^{N} \beta_i x_i - \text{MÖS}^{\text{MÖS}}^{(j)} \right)^2 \right\} \\
&= \arg\min_{\beta} \left\{ \sum_{j=1}^{K} (x_j^T \beta - \text{MÖS}^{\text{MÖS}}^{(j)})^2 \right\}.
\end{align*}
\]  \hspace{1cm} (7)

which in matrix form, is given by:
\[
\hat{\beta} = \arg\min_{\beta} \left\{ (Y - X\beta)^T (Y - X\beta) \right\}.
\]  \hspace{1cm} (8)
can be computed according to:

\[
X = \begin{bmatrix}
1 & x_1^{(1)} & \ldots & x_N^{(1)} \\
1 & x_1^{(2)} & \ldots & x_N^{(2)} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_1^{(K)} & \ldots & x_N^{(K)}
\end{bmatrix}, \quad \text{and} \quad Y = \begin{bmatrix}
\text{MOS}^{(1)} \\
\text{MOS}^{(2)} \\
\vdots \\
\text{MOS}^{(K)}
\end{bmatrix}.
\]

\(X\) 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 \(Y\) is a vector with the true MOS values. Thus, the least squares solution for \(\hat{\beta}\) can be computed according to:

\[
\hat{\beta} = (X^T X)^{-1} X^T Y
\]

C. Principal component analysis (PCA)

In order to reduce the model dimensionality without sacrificing the model accuracy, the method based on PCA was applied. This method has the goal of reducing the number of features used to estimate the MOS without losing the main information and consequently without losing the model’s accuracy.

The main difference between this section and the previous one is the fact that in addition to the high complexity model (6), the PCA is applied in order to reduce the correlation between the features used to estimate the MOS. After the application of the PCA to the group of features described in section III-A, the same strategy taken in previous section to compute the regression weights was followed.

IV. RESULTS ANALYSIS

In this section, the results obtained with the implementation of the two approaches, the low and the high complexity models, described in section III-A, are presented for both H.264 and MPEG-2 encoding standards. Two sets of data are used for training and testing. The training set is employed for model calibration, while the test set is used for evaluating the model’s accuracy. Thus, the 32 encoded video sequences evaluated in the subjective experiments have been randomly divided according to 15 sequences for training and 17 sequences for test. In what concerns to the training and testing sets selection, there were no additional criteria for selecting the video sequences for each set. In order to check for variability in the results, several combinations of training/validation set sequences have been used. Figures 14, 15 and 16 depict the results for two of these configurations. As can be observed from the figures, MOS estimates are close to their true values.

In order to validate the objective quality metric proposed in this paper, i.e., to evaluate how well the objective model predicts the subjective judgements, the model performance has been evaluated using the set of measurements proposed by the Video Quality Experts Group (VQEG) [10]. These measurements are usually addressed to as prediction accuracy, monotonicity and consistency, which are very important topics in what concern to the legitimacy of the model. They were computed using the Pearson’s correlation coefficient \((P_c)\), the Spearman’s rank order coefficient \((S_c)\) and the outlier ratio, respectively. Additionally, the root mean square error (RMS) between MOS and MOS was also measured. The results are presented in table 3. It is important to remark that when the PCA is applied, there should be a compromise between the number of reduced features and the model performance results. Thus, there is an ideal number of reduced features for which the performance is maximized, and as consequence, the number of features will be lower than the ones used in the original model (model before applying the PCA), as well as the model performance results can be similar to those presented by the original model.

At the end, after the correlation between features has been reduced, the number of features used to estimate the MOS, corresponds to the ones that have more relevance to the prediction. In this case, after the application of the PCA to the group of 7 features described in section III-A, it was possible to estimate the MOS without affecting the model accuracy with 5 and 6 features for each one of the encoding standards, H.264 and MPEG-2, respectively, as it is possible to observe from table 4.

Table 3. Performance evaluation of the proposed metrics for both compression standards, H.264 and MPEG-2.

<table>
<thead>
<tr>
<th>Performance measurement</th>
<th>Low complexity model H.264</th>
<th>Low complexity model MPEG-2</th>
<th>High complexity model H.264</th>
<th>High complexity model MPEG-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P_c)</td>
<td>0.867</td>
<td>0.931</td>
<td>0.963</td>
<td>0.974</td>
</tr>
<tr>
<td>(S_c)</td>
<td>0.916</td>
<td>0.937</td>
<td>0.957</td>
<td>0.984</td>
</tr>
<tr>
<td>Outlier ratio</td>
<td>0.250</td>
<td>0.125</td>
<td>0.031</td>
<td>0.000</td>
</tr>
<tr>
<td>RMS</td>
<td>0.666</td>
<td>0.519</td>
<td>0.370</td>
<td>0.313</td>
</tr>
</tbody>
</table>
Figure 14. Low complexity model MOS estimation results for:
(a) H.264; (b) MPEG-2.

Figure 15. High complexity model MOS estimation results for: (a) H.264; (b) MPEG-2.

Table 4. High complexity model performance after applying the PCA for H.264 and MPEG-2.

<table>
<thead>
<tr>
<th>Performance measurement</th>
<th>High complexity model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H.264</td>
</tr>
<tr>
<td>$P_r$</td>
<td>0.958</td>
</tr>
<tr>
<td>$S_c$</td>
<td>0.954</td>
</tr>
<tr>
<td>Outlier ratio</td>
<td>0.063</td>
</tr>
<tr>
<td>RMS</td>
<td>0.383</td>
</tr>
</tbody>
</table>

Figure 16. High complexity model MOS estimation results after applying PCA for: (a) H.264; (b) MPEG-2.

V. CONCLUSIONS AND FUTURE WORK

The aim of the work reported in this paper was to develop an objective model capable of predicting the MOS of compressed video sequences based only on NR features, i.e., features available at the receiver side.

In order to develop a MOS prediction model approaching the behavior of human visual system in video quality evaluation, subjective tests data were required to calibrate and validate the model. Since the majority of subjective results (e.g. those produced in MPEG groups) are only available for a restrict group of persons, this paper built its own database. The production of this database of video sequences and associated MOS, wins a new dimension of importance since the subjective results as well as of all type of information related with them, can be used in future works by people who has interest in the video quality evaluation field. In what concerns the objective quality evaluation, two new objective video quality assessment metrics were proposed. These models combine a small set of features extracted from video sequences available at the user side, in order to predict the MOS given by the observers during the subjective tests. The first considered approach (the low complexity model) did not include the MSE metric. The second approach (the high complexity model) also includes the MSE metric, which is estimated without the need of the reference video [BQ08], in order to maintain the NR property.
The subjective tests and the objective quality evaluation were conducted using two different compression standards, the MPEG-2 and the H.264/AVC.

The models’ ability to predict subjective assessment of video quality was quantitatively evaluated using three measures: the prediction accuracy (Pearson coefficient), the prediction monotonicity (Spearman coefficient) and the prediction consistency (Outlier Ratio coefficient). Furthermore, the RMS was computed in order to provide a better perception of the MOS error estimation. Based on the model performance results presented in section IV, it is possible to conclude that the two approaches are capable of correctly modeling the human visual system in video quality evaluation. However, the high complexity model shows to be the more accurate, due to the inclusion of the MSE feature. Although this second approach has the drawback of a higher computational complexity than the first one, this strategy is justified since it will improve the model accuracy.

These results confirm the good performance of the algorithm proposed in this paper. When compared with the performance of [4] and [5], where a reduced and a no-reference metrics 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).

Taking into account the model performance indicators’ values achieved for the low and high complexity models, the compression standard that shows a better MOS estimation accuracy was the MPEG-2. In order to simplify the MOS prediction model by reducing the number of features used to estimate the MOS and, as a consequence, by removing the redundancy among features, a method based on PCA was conducted. It was verified that the model accuracy is not affected, although in this case the number of features to estimate the MOS were inferior to the number of features used by the original model. For future work and in order to enhance the MOS prediction model proposed in this project a spatio-temporal model of the human visual system could be explicitly taken into account (for instance, as an explicit weight of the MSE measure). Another possible enhancement could be made in the regression model chosen to estimate the MOS, if better adjusted to each feature.

REFERENCES