From PSNR to perceived quality in H.264 encoded video sequences

Tomás Brandão  
Instituto de Telecomunicações  
ISCTE-IUL, Lisbon, Portugal 
tomas.brandao@iscte.pt

Miguel Chin  
IST-UTL, Lisbon, Portugal 
mchin@sapo.pt

Maria Paula Queluz  
Instituto de Telecomunicações  
IST-UTL, Lisbon, Portugal 
paula.queluz@lx.it.pt

ABSTRACT
This paper describes and compares a set of no-reference quality assessment algorithms for H.264/AVC encoded video sequences. These algorithms have in common a module that estimates the error due to lossy encoding of the video signals, using only information available on the compressed bitstream. In order to obtain perceived quality scores from the estimated error, three methods are presented: i) to weight the error estimates according to a perceptual model; ii) to linearly combine the mean squared error (MSE) estimates with additional video features; iii) to use MSE estimates as the input of a logistic function. The performances of the algorithms are evaluated using cross-validation procedures and the one showing the best performance is also in a preliminary study of quality assessment in the presence of transmission losses.

Keywords
No-reference, Video quality, H.264, Video transmission

1. INTRODUCTION
In order to enable new video delivery possibilities, such as QoE automatic monitoring, QoE-oriented bandwidth allocation, or even scalable billing schemes, it is desirable to have an efficient video quality assessment system that is able to compute quality scores without the need of the original signals. Objective quality metrics that belong to this class are the so called no-reference (NR) quality assessment metrics.

In this paper, three different NR metrics are described and evaluated. All of them have in common a module, proposed in [2], that estimates the quantization error due to H.264/AVC encoding of video sequences. Following this module, each metric combines the estimated error with additional information extracted from the bitstream or from the decoded video, in order to obtain the perceived quality scores.

The algorithms are evaluated using subjective data and following a cross-validation procedure. The metric showing the best performance was used in a preliminary study of video quality assessment under transmission losses (e.g. packet losses in IP networks).

The paper is organized as follows: after the introduction, section 2 provides a brief review of the error estimation module and describes the video quality prediction algorithms considered in this paper. Section 3 presents the assessment results of each algorithm. Main conclusions are given in section 4.

2. QUALITY PREDICTION MODELS
2.1 No-reference PSNR estimation
As already mentioned, all video quality metrics described in this paper use the error estimation module proposed in [2]. It provides estimates for the quantization error between an original video sequence and its H.264 encoded version. The error estimation module uses elements extracted from the video bitstream, namely the quantized DCT coefficients, \(X_k\), and their corresponding quantization steps, \(q_k\). Using these values, it computes an estimate for the squared error between the original and quantized DCT coefficient values, \(\hat{\varepsilon}^2_k\).

In order to provide a no-reference mean squared error (MSE) or PSNR prediction for the encoded video sequence, those estimates can be used in the same way as if the true error values were available:

\[
\text{PSNR}_{\text{est}}[\text{dB}] = 10 \log_{10} \frac{255^2}{\text{MSE}_{\text{est}}} \quad \text{MSE}_{\text{est}} = \frac{1}{N} \sum_{k=1}^{N} \hat{\varepsilon}^2_k ,
\]

where \(N\) is the number of considered DCT coefficients. The error can be estimated by observing the value of the quantized DCT coefficient value, \(X_k\), according to:

\[
\hat{\varepsilon}^2_k = \int_{-\infty}^{\infty} \hat{f}_X(x|X_k)(X_k - x)^2 dx,
\]

where \(\hat{f}_X(x|X_k)\) is an approximation of the statistical distribution of the original DCT coefficients values conditioned to the observed value of \(X_k\). Using Bayes rule for conditional densities, (2) can be rewritten as [2]:

\[
\hat{\varepsilon}^2_k = \int_{-\infty}^{\infty} \frac{\hat{f}_X(x|X_k)}{a_k} \hat{f}_X(x)(X_k - x)^2 dx, \quad (3)
\]

where \(a_k\) and \(b_k\) are the quantization interval limits. These can be derived from the values of the quantized DCT coefficients and their corresponding quantization step sizes, \(q_k\). As for \(\hat{f}_X(x)\), it is computed using a method that explores the correlation between distribution parameters at...
adjacent DCT frequencies and combines it with a maximum-likelihood estimation method that also uses the values of $X_k$ and $q_k$ extracted from the encoded bitstream (the details of this procedure can be found in [2]).

### 2.2 Perceptual error weighting model

The architecture of the first metric is represented in figure 1. It comprises two main blocks: a local error estimation block, which was described in the previous section, and a perceptual model whose goal is to provide weights for those estimates, according to the visibility of the corresponding error. The perceptual model is based on Kelly-Daly’s spatio-temporal contrast sensitivity function (CSF) [8, 4]. In short, this CSF is a function of the spatial frequency, $f_s$, and the retinal velocity, $v_R$, computed as:

$$\text{CSF}(v_R, f_s) = S_0 c_2 v_R (2 \pi c_1 f_s)^2 \exp \left( -\frac{4 \pi c_1 f_s}{f_{\max}} \right),$$  \hspace{1cm} (4)

where $S$ and $f_{\max}$ are defined as:

$$S = \left( s_1 + s_2 \log \left( \frac{c_2 v_R}{3} \right)^3 \right) \text{ and } f_{\max} = \frac{p_1}{c_2 v_R + 2}.$$

Using the same settings as in [8] and [4], the constants and parameters of the CSF can be set to:

- $s_1 = 6.1; s_2 = 7.3; p_1 = 45.9; c_0 = 1.14; c_1 = 0.67; c_2 = 1.7.$

The spatial frequency, $f_s$, is computed for each DCT frequency pair $(i,j)$, and depends on the distance from the observer to the screen and the dimensions of the displayed images. The velocity on the retina plane, $v_R$, is given by the angular velocity of the object on the image plane, $v_I$, compensated by a term associated to the eye movements (see [4] for additional details). $v_I$ can be estimated by using the motion vectors and the frame rate, both extracted from the bitstream:

$$v_I = f_t \left( MV_x \alpha_x \right)^2 + \left( MV_y \alpha_y \right)^2,$$  \hspace{1cm} (5)

where $MV_x$ and $MV_y$ are the components of the motion vector along the horizontal and vertical directions, respectively, $f_t$ is the frame rate of the video sequence, $\alpha_x$ and $\alpha_y$ are the components of the observation angle.

Based on the CSF’s value, computed at each location in the block-wise DCT domain, a global distortion value, $D_f$, for each video frame is computed using L4 error pooling:

$$D_f = \sqrt{\sum_k (\hat{d}_k p_k)^4},$$  \hspace{1cm} (6)

where $p_k = \text{CSF}(v_R_k, f_{sk})$ is the result of the contrast sensitivity function computed at the $k$-th DCT coefficient location and $\hat{d}_k$ is the error estimate for that coefficient. The same pooling process is applied along the time basis in order to get a global distortion metric, $D_p$, for the entire video sequence. Finally, this global value is mapped into a normalized MOS range using a logistic function similar to what is suggested by the Video Quality Experts Group (VQEG) in [10]:

$$\text{MOS}_p = a_0 + \frac{a_1}{1 + e^{a_2 + a_3 D_p}},$$  \hspace{1cm} (7)

where $a_0$ to $a_3$ are parameters that can be obtained through curve fitting.

### 2.3 Linear model

The second metric, whose architecture is represented in figure 2, is based on the work proposed in [3]. MOS predictions are computed as a linear combination of features extracted from both the bitstream and the decoded video. The considered features $f_i$ are:

- $\log(B_r)$ – the logarithm of the encoded video’s bitrate.
- $\text{MSE}$ – an estimation of the mean squared error, computed using the algorithm described in section 2.1.
- $S_I$ and $T_I$ – spatial and temporal activity values, computed using the methods recommended in [6], but using the decoded video sequences instead of the original ones.

These features are linearly combined, resulting in a MOS estimate for the received video, $\text{MOS}_p$, according to:

$$\text{MOS}_p = w_0 + \sum_{i=1}^{N} w_i f_i = f^T w,$$  \hspace{1cm} (8)

with $f = [f_1 \ldots f_N]^T$ and $w = [w_0 w_1 \ldots w_N]^T$.

In (8), $f_i$ is the value of the $i$-th feature, $w_i$ is the corresponding linear weight and $N$ is the number of features. The weights vector, $w$, can be obtained from a training set using linear regression:

$$\hat{w} = (F^T F)^{-1} F^T M,$$  \hspace{1cm} (9)

with $F = \begin{bmatrix} 1 & f^{(1)} & \ldots & f^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & f^{(K)} & \ldots & f^{(K)} \end{bmatrix}$, and $M = \begin{bmatrix} \text{MOS}^{(1)} \\ \vdots \\ \text{MOS}^{(K)} \end{bmatrix}$.

$F$ is a $K \times N$ matrix, where each row contains the feature values extracted from the $k$-th degraded video sequence in the training set and $M$ is a vector with the true MOS values of the sequences included in the training set.

### 2.4 Logistic model

The last model, whose architecture is represented in figure 3, is an hybrid model that combines the MSE estimate produced by the error estimation block with a spatio-temporal activity index computed from the decoded video.

The spatio-temporal activity index, $s_i$, is defined as the maximum between the spatial activity and temporal activity, computed as [1]:

![Figure 1: Perceptual weighting model scheme.](image1.png)

![Figure 2: Linear model scheme.](image2.png)
3. RESULTS

3.1 Subjective data

In order to evaluate the performance of the models described in the previous section, the subjective quality assessment data available at the IT-Image Group site\(^1\) has been used. This subjective data has been obtained by carrying out a set of double stimulus subjective tests that followed the Degradation Category Rating (DCR) methodology, described in Recommendation ITU-T P.910 [6].

The database comprises 12 reference video sequences that span a large range of spatio-temporal activities. The test conditions considered during the subjective quality assessment sessions are versions of the reference sequences encoded at different bitrates (from 32 kbit/s to 2 Mbit/s) using the JM 12.4 H.264/AVC codec [5]. The total number of test conditions in this database is 58, each one assessed by at least 18 validated subjects.

3.2 Quality prediction results

Since all the algorithms presented in the paper require parameter estimation procedures, a cross-validation training methodology has been followed. The adopted procedure resembles the leave-one-out cross-validation method [9]: the subjective database has been organized according to 12 “families”, where each “family” is the set of H.264 encoded versions of each reference video sequence, together with the corresponding MOS values. The training/validation process is performed by turns. In each turn, one family is used as the validation set and the remaining 11 families are used as the training set. The process is repeated until every family takes its place in the validation set.

The predicted MOS values that result from each validation turn are depicted in figures 6-a) to d) and confronted with their true values. Figures 6-a) to c) are the outcomes of the three algorithms described in the paper, while figure 6-d) depicts the result of MOS predictions using the PSNR only. It can be observed that the logistic model described in section 2.4 holds the best results, leading to MOS predictions that are quite close to the ground truth data.

In [10], VQEG suggests a set of statistical measurements whose goal is to evaluate the performance of an objective metric. These measurements are the root mean squared error (RMS), the Pearson’s correlation coefficient (CC), the Spearman’s rank order coefficient (RC) and the outlier’s ratio (OR). Using these indicators, a comparison between the model’s performances can be observed in table 1. It confirms that the logistic model leads to the best results. The perceptual error weighting algorithm holds values of CC and RC above 0.90, but its RMS and OR results are strongly penalized due to about 5 points whose MOS predictions are far from their true values. The linear model algorithm seems to be the weakest. Nevertheless, all models lead to results that are substantially better than using the PSNR.

3.3 MOS prediction under transmission losses

A preliminary study on the applicability of the described

\(^1\)http://amalia.img.lx.it.pt/~tgab/H264_test
models to transmission scenarios with packet losses has also been performed. The subjective quality database from Politecnico di Milano has also been used in a preliminary study. While the results under transmission losses seem to be promising, no solid conclusions can be drawn yet due to the limited amount of subjective data available for these experiments.

5. REFERENCES


\[ MOS_p = MOS_{p0} \exp \left( \frac{PL}{\theta} \right), \]

where $MOS_{p0}$ is an initial prediction for MOS, without considering the effect of the missing packets (i.e., computed by any of the presented metrics, using the information that effectively arrives at the decoder), $PL$ is the packet loss rate, and the parameter $\theta$ is a constant (obtained using data regression). Note that this model resembles the model proposed in ITU-T Rec. G.1070 [7].

Figure 7-b) depicts MOS vs. $MOS_p$ results, using the subjective quality data mentioned above and the logistic model presented in section 2.4, for computing $MOS_{p0}$ in (11). Following a procedure similar to the presented in 3.2, the value of $\theta$ was obtained using leave-one-out cross-validation. The performance indicators achieved on this test were RMS=0.48, CC=0.94 and RC=0.93.

4. CONCLUSIONS

Three different no-reference video quality assessment algorithms have been described and evaluated. Those algorithms share a common component on their architecture – they all use an algorithm that computes an estimation for the error due to lossy video encoding. The algorithms’ performances have been evaluated using a cross-calibration procedure over 58 subjective test conditions (which are H.264 encoded versions of 12 different video sequences). The results have shown that the logistic model algorithm has the best performance.

In order to assess the quality of video sequences subject to packet losses on IP networks, the logistic model algorithm has also been used in a preliminary study. While the results under transmission losses seem to be promising, no solid conclusions can be drawn yet due to the limited amount of subjective data available for these experiments.

![Figure 6: MOS prediction results.](image1)

![Figure 7: MOS prediction for transmission losses.](image2)

![Table 1: Evaluation of the described metrics.](table1)

<table>
<thead>
<tr>
<th>Metric</th>
<th>RMS</th>
<th>CC</th>
<th>RC</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptual error weighting</td>
<td>0.52</td>
<td>0.91</td>
<td>0.92</td>
<td>0.14</td>
</tr>
<tr>
<td>Linear model</td>
<td>0.64</td>
<td>0.86</td>
<td>0.86</td>
<td>0.16</td>
</tr>
<tr>
<td>Logistic model</td>
<td>0.32</td>
<td>0.97</td>
<td>0.96</td>
<td>0.05</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.83</td>
<td>0.75</td>
<td>0.76</td>
<td>0.28</td>
</tr>
</tbody>
</table>

\[ MOS_{p} = MOS_{p0} \exp \left( \frac{PL}{\theta} \right), \]