NO-REFERENCE PERCEPTUAL QUALITY METRIC FOR H.264/AVC ENCODED VIDEO

Tomás Brandão[*,†] *and Maria Paula Queluz*[*,‡]

[*] Instituto de Telecomunicações; [†] ISCTE – Lisbon University Institute;
[‡] Instituto Superior Técnico – Technical University of Lisbon, Portugal.
Emails: {tomas.brandao, paula.queluz}@lx.it.pt

ABSTRACT

This paper extends a no-reference PSNR algorithm, previously derived by the authors, by incorporating a spatio-temporal model of the human visual system (HVS). Video errors due the H.264/AVC encoding process are firstly estimated using the received DCT coefficients and the corresponding quantization steps. In order to obtain a metric for the visible video distortion, the estimated errors are then weighted using an HVS model based on the spatio-temporal contrast sensitivity function derived by Kelly and Daly. The video related inputs for the perceptual model are the motion vectors and the frame rate, which are also extracted from the received encoded video. In order to validate the results, a set of video sequences that span a wide range of content have been encoded at different bitrates and their quality has been subjectively assessed. Results show that the quality scores computed by the proposed metric are well correlated with the mean opinion scores resulting from the subjective assessment.

Index Terms— Video quality, no-reference metric, perceptual model, H.264

1. INTRODUCTION

Quality assessment of multimedia data has become an important matter, especially due to the increasing transmission of video contents over the internet and mobile networks.

The most reliable source for assessing the quality of multimedia are the human viewers since they are the consumers of video communications products. However, gathering quality assessment data from the human viewers requires the completion of *subjective quality tests*, which must be conducted under controlled test conditions. Thus, subjective quality scores cannot be used in real-time applications.

An alternative is to score video quality using *objective metrics*. Ideally, an objective metric should compute quality scores matching the subjective ones. Most of the research performed on this field has been focused on *full reference* (FR) metrics (some examples in [1–3]), which require both the original and the distorted media to compute the quality scores. However, this class of metrics is not suitable for media distribution scenarios, since the original data is usually unavailable at the receiver.

Thus, in transmission environments, a quality measurement system at the receiver should be able to provide quality feedback without requiring the reference signals. This had lead to an increased research effort on *no-reference* (NR) quality metrics (*e.g.* [4]) and *reduced reference* (RR) quality metrics (*e.g.* [5]). NR metrics rely on the received signals only, while RR metrics use the received signals and a certain amount of information about the reference signal (sent through a side information channel).

The method proposed in this paper belongs to the NR quality metrics category. It computes the quality scores of H.264 encoded video sequences based on the quantized *discrete cosine transform* (DCT) coefficients, on their corresponding quantization steps and on the motion vectors, which can all be extracted from the received bitstream. Basically, it consists of a local error estimation procedure followed by a perceptual weighting of the resulting estimates.

Local error estimation is performed based on the *peak signal to noise ratio* (PSNR) estimation algorithm proposed in [6], which relies on statistical properties of the block-based DCT coefficient data. In short, coefficient's distributions are modeled according to Cauchy or Laplace *probability density functions* (PDFs). The parameters of those PDFs are computed based on the received quantized DCT coefficients, using a maximum-likelihood parameter estimation method combined with a linear prediction. This prediction scheme explores the correlation between PDF parameter values located at neighbor DCT frequencies.

Error weighting is performed using a spatio-temporal perceptual model based on the work by Kelly and Daly. In [7], Kelly devised an analytic model for the spatio-temporal *contrast sensitivity function* (CSF), based on data collected from his experiments. His work is extended by Daly in [8], by considering movements of the eye, namely *smooth pursuit*, *natural drift* and *saccadic* eye movements.

In order to evaluate the results of the proposed metric, subjective tests have been conducted in accordance with Recommendation ITU-T P.910 [9]. A set of representative video sequences have been encoded at different bit rates with the

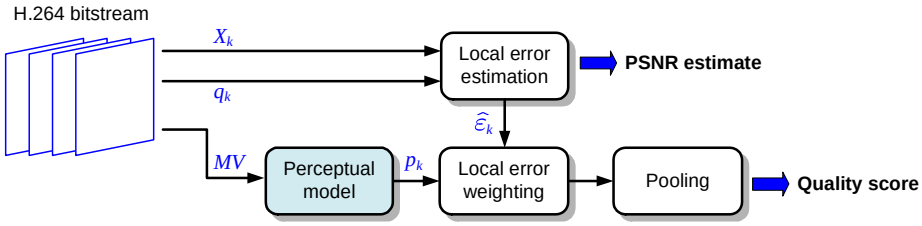


Fig. 1. Proposed quality assessment metric.

H.264/AVC standard. The resulting *Mean Opinion Scores* (MOS) constitute the benchmark relatively to which the performance of the proposed algorithm is evaluated. Quality predictions of the proposed algorithm have shown a good correlation with the MOS values resulting from the subjective quality assessment tests.

This paper is organized as follows: section 2 describes the architecture of the proposed no-reference quality assessment metric, emphasizing the perceptual model. Section 3 describes the methodology and conditions associated to subjective tests. Results are depicted in section 4 and finally, the main conclusions and topics for further research are given in section 5.

2. QUALITY ASSESSMENT METRIC

2.1. General architecture

The proposed architecture for assessing the quality of an H.264 encoded video sequence is represented in figure 1. It consists of two main steps: local error estimation and perceptual weighting of those estimates. This type of architecture allows to extend the use of perceptual models (based on error weighting) to the no-reference quality assessment problem.

The error estimation module is based on the algorithm proposed in [6], whose inputs are the quantized DCT coefficients X_k and the corresponding quantization steps q_k. Using these values, it computes an estimate for the squared error between the original and quantized DCT coefficient values, $\hat{\varepsilon}_k^2$. At this point, the algorithm is able to estimate the PSNR of the received sequence, using squared error estimates instead of their true values:

$$\mathrm{PSNR}_{\mathrm{est}_{[\mathrm{dB}]}} = 10\log_{10}\frac{255^2}{\frac{1}{N}\sum_{k=1}^{N}\hat{\varepsilon}_k^2}, \qquad (1)$$

where N is the number of DCT coefficients. Note that, in accordance with Parceval's theorem, it is indifferent to measure the PSNR in the pixel or in the DCT domain. The error estimates are then perceptually weighted, using a spatio-temporal model based on [7, 8]. The function of this model is to compute local perceptual weights p_k. The inputs for the model are the motion vectors, MV, and the video frame rate, f_r, both extracted from the encoded bitstream. The weights p_k and error estimates $\hat{\varepsilon}_k$ are then combined and pooled in order to obtain a global value of the distortion for the whole frame (or for a sequence of frames).

2.2. Spatio-temporal CSF model

It his known that the human visual system is more sensitive to image contrast rather than the absolute luminance values. Contrast can be defined as the ratio between the local luminance variation and the average background luminance. *Contrast sensitivity* is the inverse of the minimum contrast necessary for an observer to detect a stimulus. A spatio-temporal CSF describes the evolution of the HVS sensitivity to luminance changes and it depends on the spatial and temporal frequencies of the stimulus. Based on data collected from his psychophysical experiments [7], Kelly proposed a spatio-temporal CSF model as a function of the spatial frequency, f_s, and the retinal velocity, v_R, which implicitly gives the temporal frequency. This function is given by:

$$\mathrm{CSF}(v_R, f_s) = Sc_0c_2v_r(2\pi c_1 f_s)^2 \exp\left(-\frac{4\pi c_1 f_s}{f_{max}}\right), \qquad (2)$$

with the terms S and f_{max} defined as:

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

The constants s_1, s_2 and p_1 have been set to 6.1, 7.3 and 45.9, respectively [7]. The parameters c_0, c_1 and c_2 allow model tunning and have been set to the same values as in [8]: $c_0 = 1.14$, $c_1 = 0.67$ and $c_2 = 1.7$.

The spatial frequency f_s can be computed as the euclidean norm of the subband spatial frequency components:

$$f_s = \sqrt{f_x^2 + f_y^2}. \qquad (3)$$

In the $K \times K$ block-wise DCT domain, the components f_y and f_x of the spatial subband frequency (in cycles per degree) at location (i, j) of a DCT block are given by:

$$f_y = \frac{i}{2K\alpha_y} \text{ and } f_x = \frac{j}{2K\alpha_x}, \qquad (4)$$

where α_x and α_y the observation angle of a pixel along the horizontal and vertical directions, respectively. The the observation angle of a pixel along a generic direction ϕ can be

computed as:

$$\alpha_\phi = \arctan \frac{l_\phi}{2dN_\phi} \simeq \frac{l_\phi}{2dN_\phi}. \quad (5)$$

where l_ϕ is the height/width of the images displayed on the screen, d is the distance from the observer to the screen and N_ϕ is the vertical/horizontal resolution of the displayed video sequence.

The object velocity on the retina plane is strongly related with the object velocity in the image plane. However, the human eye has the ability to track objects, slowing down the velocity of the object in the retina plane. This characteristic is called the *smooth pursuit eye movement* (SPEM). Additionally, there are other movements of the eye, namely the *natural drift* and *saccadic* eye movements. The former is a slow eye movement that causes a little amount of motion in the retina plane, while the latter are fast eye movements cause by changing the gaze to new image plane locations.

According to [8], the retinal image velocity can be computed as:

$$v_R = v_I - v_E, \quad (6)$$

where v_I is the angular velocity of the object on the image plane and v_E is a compensation term associated to the eye movements, computed as:

$$v_E = \min\{g_S \times v_I + v_{MIN}; v_{MAX}\}, \quad (7)$$

where g_S is the SPEM gain, set to 0.92; v_{MIN} and v_{MAX} are the minimum and maximum velocities associated to the eye natural drift and saccadic eye movements, set to 0.15 and 80 deg/s, respectively.

The angular velocity on the image plane, v_I, is given by:

$$v_I = f_r \sqrt{(MV_x \alpha_x)^2 + (MV_y \alpha_y)^2}, \quad (8)$$

where MV_x and MV_y are the components of the motion vector along the horizontal and vertical directions, respectively, and f_r is the frame rate of the video sequence. The components of the observation angle of a pixel, α_x and α_y, are those resulting from (4).

2.3. Quality scores

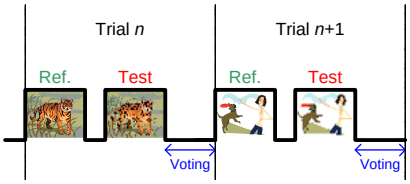
Based on the result of the CSF computed at each location in the block-wise DCT domain, a global distortion value for the whole video frame, D_f, is then computed using $L4$ error pooling according to [10]:

$$D_f = \sqrt[4]{\sum_k (\hat{\varepsilon}_k p_k)^4}, \quad (9)$$

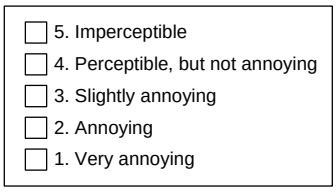
where $p_k = \text{CSF}(v_{R_k}, f_{s_k})$ is the result of the contrast sensitivity function computed at the k-th DCT coefficient location and $\hat{\varepsilon}_k$ is the local error estimate computed by the error estimation block. Finally, the same pooling process is applied along the time basis in order to get a global distortion metric for the encoded video sequence:

$$D_g = \sqrt[4]{\sum_i {D_{f_i}}^4}. \quad (10)$$

Note that, for longer video sequences, a granularity period for computing D_g could be defined (*e.g.*, D_g could be computed every 10 seconds of video).



(a) Test session structure.



(b) Impairment scale.

Fig. 2. Degradation Category Rating Methodology.

3. SUBJECTIVE QUALITY ASSESSMENT

3.1. Methodology

The subjective quality assessment tests have been performed in accordance with Recommendation ITU-T P.910 [9]. The method followed in this work was the *Degradation Category Rating* (DCR) [9] (which corresponds to the *Double Stimulus Impairment Scale* (DSIS) method in [11]). In this methodology, the observer is presented with video sequences organized in pairs, as illustrated in figure 2-a): the first to be displayed is called the *reference* sequence (usually, the original) while the second is called the *test* or *impaired* sequence (in our case, the result of lossy encoding); it has been used the five grade impairment scale depicted in figure 2-b), that reflects the observer´s judgment about the impairment.

3.2. Assessment conditions

According to [9], at least 15 observers are needed in order to produce reliable results. In our case, 22 observers (mostly students) participated in the subjective experiments. They were screened for visual acuity and color blindness, using a Snellen Eye Chart and Ishihara's plates, respectively. The duration of

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 inactive screen to peak luminance	0.023
Ratio of background to peak luminance	0.14

Table 1. Environmental viewing conditions.

Sequence	**Bit rates**
Coastguard	66, 131, 263 and 525 kbit/s
Container	66, 131, 262 and 524 kbit/s
Football	264, 526, 1051 and 2105 kbit/s
Foreman	131, 263, 525 and 1051 kbit/s
Mobile	131, 262, 525 and 1049 kbit/s
Stephan	140, 263, 525 and 1050 kbit/s
Table	66, 132, 263 and 525 kbit/s

Table 2. Bit rates of the sequences used in the tests.

each session was about 20 minutes with the room setup allowing two observers to participate in each session.

As for the environmental viewing conditions, three factors must be considered: the lighting, the ambiance noise and the quality and calibration of the display. Two high quality LCD displays of the same model with native resolutions of 1650×1050 pixels have been used and they were previously calibrated. The display and room characteristics used in the subjective tests, listed in table 1, are within the values recommended in [9].

3.3. Selection of test material

In order to avoid boring the observers and get meaningful results, it is important that a small, but representative, set of video sequences is used during the tests. In particular, the spatial and temporal activities are important parameters which should be considered when choosing the test sequences. Thus, it is important to choose a set of sequences that span a wide range of values for those activities. The literature provides several methods of measuring a video spatial and temporal activity. In this work, the methods recommended in [9] have been used:

- **Spatial activity**: the horizontal and vertical picture gradient are computed using the well known Sobel filters. The gradient norm (the square root of the sum of the vertical and horizontal gradient squares) is then computed for each pixel. The standard deviation of the gradient norm is calculated for each frame, resulting in a time series of frame-by-frame spatial activities. 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 is obtained by computing the difference, pixel-by-pixel, between each pair of successive frames. 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.

Due to changes of the camera perspective during video acquisition or scene cuts, the global activity measurements could have a high value even if the sequence has a low temporal and/or spatial activity. In order to minimize this effect, the global activity values result from applying the 95% percentile to the temporal and spatial activities series, instead of using its maximum.

Figure 3 represents the video sequences used in the subjective tests. They have been selected based on their spatio-temporal activities, whose values are depicted in figure 4. These sequences are CIF format (352×288), with a frame rate of 30 Hz. The sequences were encoded using the reference H.264 software tools [12]. For each sequence, four different bit rates in the range from 32 to 2048 kbit/s have been used for encoding, resulting in the bitrates summarized in table 2. GOP-15 structure $IBBPBBP...$ has been used in all encoding runs. Only the 4×4 transform size was allowed and the low complexity rate-distortion optimization algorithm provided on the software has been used. The result is a set of 28 encoded sequences (impaired sequences), whose qualities were judged by the test participants. This set allows to test the human visual system (HVS) perception to different kinds of video qualities and to indirectly force the observers to use all grades of rating scale.

3.4. 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 [11]-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.

The resulting MOS values and the video sequences used in the subjective quality assessment tests are available online at `http://amalia.img.lx.it.pt/~tgsb/H264_test/`.

4. RESULTS

The input video sequences used for evaluation of the proposed quality metric are the same used in the subjective quality assessment tests, represented in figure 3.

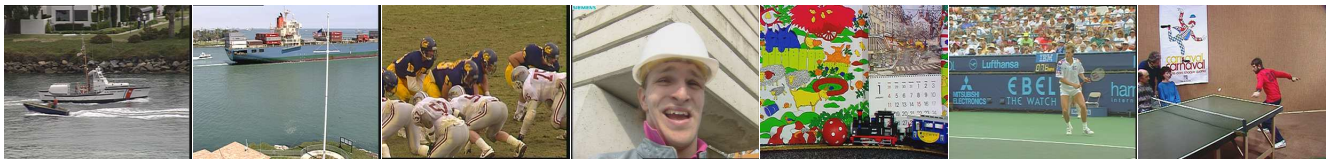


Fig. 3. Video sequences selected for the subjective tests. From left to right: *Coastguard*; *Container*; *Football*; *Foreman*; *Mobile & Calendar*; *Stephan*; *Table-tennis*.

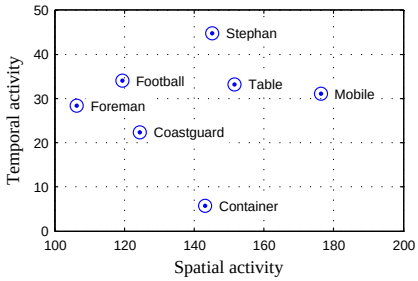


Fig. 4. Spatio-temporal activities of the selected sequences.

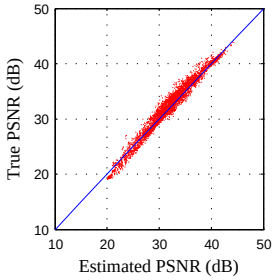


Fig. 5. PSNR estimates *vs.* their true values.

4.1. PSNR estimation

The PSNR has been estimated and compared with its true value. Results are depicted in figure 5. As can be observed from the plots, the proposed method is quite accurate. Additionally, table 3 evaluates these results: the symbols ε_{avg}, ε_{rms} and ρ stand for the average error, the root mean square error and the correlation between true and estimated PSNR values, respectively. As can be observed from the table, the PSNR estimation method is quite accurate regardless of the frame type.

4.2. Quality scores

The results for quality assessment have been evaluated by comparing the quality scores retrieved by the algorithm with the ones that result from the subjective experiments described in section 3.

Figure 6-a) depicts the the value of the propose perceptual distortion metric given by equation (10) versus the corresponding MOS values. Following a procedure similar to what is suggest by the *Video Quality Experts Group* (VQEG) in [13], a logistic function was used in order to map the D_g values into the MOS range used in the experiments:

Frame Type	ε_{avg}	ε_{rms}	ρ
I	0.69	0.86	0.99
P	0.83	1.09	0.98
B	0.89	1.13	0.98
All	0.86	1.10	0.98

Table 3. PSNR estimation error statistics.

$$\text{Estimated MOS } = a_0 + \frac{a_1}{1 + e^{a_2 + a_3 D_g}}, \qquad (11)$$

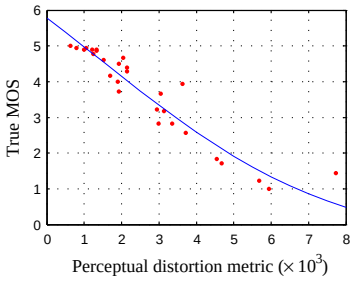
where a_0 to a_3 are the curve fitting parameters. These parameters have been computed in order to minimize the square differences between the estimated MOS scores their true MOS values and have been computed using the *Levenberg-Marquardt* method for non-linear least squares minimization problems. The resulting logistic function is also plotted in figure 6-a). Figure 6-b) shows the quality scores that result from the proposed algorithm versus the MOS values obtained in the subjective tests. As can be observed, the NR objective quality scores resulting from the proposed algorithm are well correlated with the subjective quality assessment data.

In [13], VQEG suggests a set of statistical measurements in order to benchmark the performance of an objective metric. These indicators can be observed in table 4 and confirm the that the proposed NR quality metric performs well.

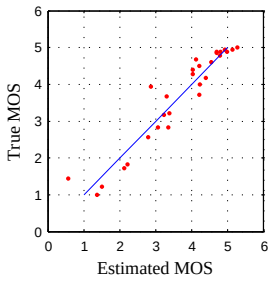
When comparing these results with other results found on the literature, the proposed method seems to outperform other algorithms designed for similar purposes. In [4], Ries *et al.* propose a no-reference video quality assessment metric where the video contents are classified, and quality scores result from combining a set of motion features. The declared performance in [4] is CC $= 0.86$, which is below the results of the method proposed in this paper.

In [5], Oelbaum and Diepold propose a reduced reference method for H.264 encoded sequences where several features extracted from the video are combined (most of them are artifact measurements and motion oriented features), and the results are adjusted based on two parameter values sent through a side channel. The declared performance of this method is CC $= 0.84$, RC $= 0.80$ and OR $= 0.58$, which are also below the results achieved by the algorithm proposed in this paper.

A standard for reduced reference quality assessment of



(a) Distortion metric *vs.* true MOS.



(b) Estimated *vs.* true MOS.

Fig. 6. MOS estimation results.

Root mean square error (RMS)	0.383
Pearson correlation coefficient (CC)	0.953
Spearman rank order coefficient (RC)	0.946
Outliers ratio (OR)	0.071

Table 4. Evaluation of the proposed metric.

cable television signals is given in Recommendation ITU-T J.246 [14]. This RR metric – *Edge-PSNR* – is based on edge maps extracted from the original signals, which are sent to the receiver. The performance of this metric increases has the side channel bandwidth increases (*i.e.*, as the number of points in the sent edge map increases). The resulting values for CC are in the range $0.81 - 0.83$. Again, our method shows better performance. However, it must be kept in mind that the method proposed in this paper is oriented to H.264 encoding while the standardized method is not distortion specific.

5. CONCLUSIONS

A no-reference video quality assessment algorithm incorporating some aspects of the human visual system has been proposed. Although the H.264/AVC standard has been considered, the method could be straightly applied to other DCT-based video encoding schemes. The algorithm comprises an error estimation module followed by an error weighting module based on a spatio-temporal CSF model. The resulting MOS estimates correlate well with the human perception of quality and show better results than other algorithms found on literature.

As for future work, the algorithm should be extended in order to deal with transmission errors (*i.e.*, packet losses) and a more complete perceptual model could also improve the algorithm's performance.

6. REFERENCES

[1] S. Winkler, "A perceptual distortion metric for digital color video," in *proc. of SPIE*, vol. 3644, S. Jose, USA, 1999.

[2] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Image Communication - Special issue on Objective Video Quality Metrics*, vol. 19, no. 2, pp. 121–132, February 2004.

[3] E. Ong, X. Yang, W. Lin, Z. Lu, and S. Yao, "Perceptual quality metric for compressed videos," in *proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Philadelphia, USA, March 2005.

[4] M. Ries, O. Nemethova, and M. Rupp, "Performance evaluation of mobile video estimators," in *proc. of EUSIPCO - European Signal Processing Conference*, Poznan, Poland, September 2007.

[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.

[6] 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.

[7] D. H. Kelly, "Motion and vision II: stabilized spatio-temporal threshold surface," *Journal of the Optical Society of America*, vol. 69, no. 10, pp. 1340–1349, October 1979.

[8] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Vision model and applications to image and video processing*, Ed. C. van den Branden Lambrecht, Kluwer, 2001.

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

[10] A. B. Watson, "DCT quantization matrices optimized for individual images," in *proc. of SPIE Human Vision, Visual Processing, and Digital Display IV*, S. Jose, USA, 1993.

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

[12] Heinrich-Hertz-Institut, "JM 12.4 – H.264 reference software," December 2007, available online at `http://iphome.hhi.de/suehring/tml/`.

[13] 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.

[14] ITU-T, "Recommendation J.246 – Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference." 2008.