Objective Video Quality Metrics

M. Vranješ, S. Rimac-Drlje, D. Žagar
University of Osijek, Faculty of Electrical Engineering
Kneza Trpimira 2B, Osijek, HR-31000, Croatia
E-mail: *rimac@etfos.hr*

Abstract – *This paper discusses methods used in different objective video quality metrics. An experimental comparison of different objective methods is also conducted. This experiment shows the importance of video content for a subjective quality evaluation not comprised well by the objective metrics used.*

Keywords: *objective quality metrics, video quality, subjective testing*

1. INTRODUCTION[1]

In the last decade there has been an increasing interest in developing objective quality metrics for evaluation of different types of digital video distortion. In a heterogeneous communication environment different compression techniques and different bit rates can be used simultaneously, and different types of errors can occur as well as different demands for Quality of Service (QoS). Due to their simplicity, the mean-squared error (MSE) and the peak signal-to-noise ratio (PSNR) are very widely used quality metrics. Usually they cannot give an objective quality measure corresponding well to the quality perceived by a human observer for a wide range of coding and transmission parameters.

Several objective metrics have been recently developed showing a good correspondence with the subjective Mean Opinion Sc ore (MOS) obtained from human observers, [1]-[7]. In this paper we discuss full-reference models and give experimental results for three objective metrics compared with subjective results.

2. OBJECTIVE QUALITY METRICS

Basically, two different approaches are used in objective quality metrics. Some metrics use Human Vision System (HVS) characteristics to extract important features in the original and a distorted signal and valuate differences between them according to characteristics of human visual error sensitivity, [1]-[4]. The second approach is based on structural distortion measures (like blurring of edges or visibility of blocks), [5]-[7]. We present a Standard Spatial Observer (SSO) based model, [2], and DCT based VQM model exemplifying the first approach, and the General Video Quality Model (VQM_G), [6], and the Structural Similarity (SSIM) index, [7], as examples of the second approach.

2.1. SSO model

In [2], Watson and Malo propose a quality metrics based on the standard spatial observer model that incorporates psychophysical and physiological vision research results. The basic model uses a Contrast Sensitivity Function (CSF) in the spatial domain to extract features important for a human observer. The frame difference between the original frame p(x,y,t) and the distorted one p'(x,y,t) is d(x,y,t)

$$d(x,y,t) = p(x,y,t) - p'(x,y,t) \tag{1}$$

The frame difference is further filtered by multiplication with CSF in the frequency domain

$$d(x,y,t)_{SSO} = F_S^{-1}(CSF_S \cdot F_S(d(x,y,t))) \tag{2}$$

In (2) F_S^{-1} and F_S are the inverse Fourier transform and the Fourier transform, respectively. CSF_S is the spatial CSF function. Visibility of the frame difference, d(t), is computed by pooling d(x,y,t) over space by Minkovski summation

$$d(t) = \left(\sum_{x,y} d(x,y,t)_{SSO}^{2.9} \right)^{1/2.9} \tag{3}$$

The overall difference for a video sequence, d, is calculated by pooling d(t) over time by Minkovski summation given by

$$d = \left(\sum_{t} d(t)^2 \right)^{1/2} \tag{4}$$

The SSO model uses four different processing of $d(x,y,t)_{SSO}$ prior to spatial and temporal pooling. One is a temporal pre-filter before spatial summation

$$d(x,y,t)_{SSO+t} = F_t^{-1}(CSF_t \cdot F_t(d(x,y,t)_{SSO})) \tag{5}$$

[1] This research is supported by the Croatian Ministry of Education, Science and Sports through the project No.165-0361630-1636.

where F_t^{-1} and F_t are the inverse and the direct temporal Fourier transform, respectively, and CSF_t is the temporal CSF function.

The model also uses a temporal post-filter after spatial summation, given by

$$d(x,y,t)_{SSO+p} = F_t^{-1}(CSF_t \cdot F_t(d(t))). \quad (6)$$

A local masking model is used in the third processing of $d(x,y,t)_{SSO,}$ as given by

$$d(x,y,t)_{SSO+m} = \frac{d(x,y,t)_{SSO}}{\sqrt{1+\left(\frac{h(x,y,t)*p(x,y,t)}{c}\right)^2}} \quad (7)$$

where the Gaussian kernel $h(x,y,t)$ is convoluted with the frame $p(x,y,t)$ to achieve localization of masking. The parameter c optimizes the strength of masking.

Finally, the SSO based model uses an algorithm for detection of field duplication (which is done with some coders) and decreases overestimation of the error observed in distortion measures based on pixel-by-pixel comparisons. The SSO model with this field replication algorithm is marked as SSO+h.

Authors have tested this model on the full set of 160 VQEG, 30 Hz sequences, [2]. The highest correlation between model prediction d and subjective Difference MOS (DMOS) is obtained for the SSO model SSO+m+h with local masking and the field replication algorithm, as well as for the following models: SSO+m+p+h, SSO+t+m+p+h, SSO+t+m+h, SSO+m+p, SSO+t+m+p, SSO+t+m and SSO+m. All of these models have used local masking and that has shown that local masking as a single processing has the most significant influence on prediction accuracy.

In the same research, for the purpose of comparison, results of the MSE alone and the MSE with local masking are introduced. It is worth mentioning that MSE with masking (MSE+m) has only slightly worse prediction results than the best SSO models.

2.2. DCT- based VQM model

VQM model is a DCT-based video quality metric, developed by F. Xiao, [3]. This metric is based on simplified human spatial-temporal contrast sensitivity model. Model calculates distortion of a compressed video in four steps:

1. For every frame the model performs Discrete Cosine Transform (DCT) for 8 x 8 pixels blocks $b_i(x,y,t)$ of the original video frame p(x,y,t), and for blocks $b_i'(x,y,t)$ of the distorted video frame p'(x,y,t).

$$DCTb_i(u,v,t) = DCT(b_i(x,y,t))$$
$$DCTb_i'(u,v,t) = DCT(b_i'(x,y,t)) \quad (8)$$

2. The model converts DCT coefficients to local contrast values $LC_i(u,v,t)$ by using DC component of each block.

$$LC_i(u,v,t) = \frac{DCTb_i(u,v,t)\cdot\left(\frac{DC_i}{1024}\right)^{0.65}}{DC_i}$$
$$DC_i = DCTb_i(0,0,t) \quad (9)$$

On the same way model obtains $LC_i'(u,v,t)$ of the compressed video.

3. The model converts $LC_i(u,v,t)$ and $LC_i'(u,v,t)$ to just noticeable difference values, $JND_i(u,v,t)$ and $JND_i(u,v,t)$, by using static and dynamic spatial contrast sensitivity function (CSF).

4. The JND coefficients of original and compressed sequences are subtracted to produce a difference values $Diff_i(t)$. Model incorporates contrast masking into simple maximum operation and then weights it with the pooling mean distortion. Final VQM score is obtained by:

$$Dist_{Mean} = 1000 \cdot mean(mean(Diff_i(t)))$$
$$Dist_{Max} = 1000 \cdot \max(\max(Diff_i(t)))$$
$$VQM = Dist_{Mean} + 0.005 \cdot Dist_{Max} \quad (10)$$

The VQM score decreases as quality of compressed video rises and it is zero for the losless compressed video.

2.3. NTIA General Video Quality Model (VQM_G)

The VQM_G model is developed by the National Telecommunications and Information Administration (NTIA) and obtained the best average correlation for both 525-line video (NTSC) and 625-line video (PAL) sequences in Video Quality Experts Group (VQEG) Phase II Full Reference Television tests, [8]. Furthermore, this model has been standardized by the American National Standards Institute (ANSI), [6].

The VQM_G model was designated to be a general-purpose quality model for a wide range of video systems with different resolution, frame rates, coding techniques and bit rates. The model uses seven parameters based on different quality features of a video stream. For all features the VQM_G model performs basically the same steps. In the first stage a filter is applied to the original and a distorted video to enhance some property important for quality. After that, features $f_i(t)$ are extracted from the spatial-temporal sub-region, $S\text{-}T_i$, using the mean or standard deviation of each filtered S-T region. Finally, a quality parameter $q_i(t)$ is obtained by comparing quality features of the original video, $f_i(t)$

and features of the disturbed video, $f_i'(t)$. One of the following comparison functions is used for calculation of quality parameters:

a) Euclidean distance

$$q_i(t) = \sqrt{(f_i(t) - f_i'(t))^2 + (f_{i2}(t) - f_{i2}'(t))^2} \quad (11)$$

b) The ratio comparison function

$$q_i(t) = \frac{(f_i(t) - f_i'(t))}{f_i(t)} \quad (12)$$

c) The log comparison function

$$q_i(t) = \log_{10} \frac{f_i'(t)}{f_i(t)} \quad (13)$$

Spatial and temporal pooling is obtained by using some form of worst case processing (e.g. average of 10% worst-case $q_i(t)$ values). That makes the worst part of video the predominant feature in the quality measure. A brief description of quality parameters used in the VQM_G model is given in the remainder of this section. For more information interested readers are referred to [6].

The first model parameter is *si_loss* that detects loss of spatial information (e.g. blurring). Video frames are filtered (horizontally and vertically) with a spatial filter, which enhances the information of edges in a video frame.

Parameter *hv_loss* detects a shift of edges from horizontal and vertical to diagonal orientation.

Parameter *hv_gain* detects a shift of edges from diagonal to horizontal and vertical orientation (e.g. blocking).

The *si_gain* parameter measures improvements of quality caused by edge sharpening or enhancement.

Parameter *chroma_spread* detects changes in the spread of color samples distribution.

Parameter *chroma_extreme* detects severe localized color impairments.

Parameter *ct_ati_gain* measures perceptibility of spatial impairments in dependence of the amount of motion as well as perceptibility of temporal impairments in dependence of the amount of spatial data.

The VQM_G measure consists of a linear combination of the described parameters:

$$\begin{aligned} VQM_G = {} & -0.2097 \cdot si\_loss + 0.5969 \cdot hv\_loss + 0.2483 \cdot \\ & \cdot hv\_gain + 0.0192 \cdot chroma\_spread - \\ & - 2.3416 \cdot si\_gain + 0.0431 \cdot ct\_ati\_gain + \\ & + 0.0076 \cdot chroma\_extreme \end{aligned} \quad (14)$$

For no perceived impairment the model gives the output value equal to zero, and for a rising level of impairment the output value rises, too.

The VQM_G model participated in VQEG tests, which include 1,536 subjectively rated video sequences. The Pearson linear correlation between VQM_G and subjective DMOS was 0.938 for 525-line test sequences, and 0.886 for 625-line sequences.

In the same test, the PSNR measure obtained Pearson correlation 0.804 for 525-line test sequences, and 0.733 for 625-line test sequences.

2.4. Structural Similarity (SSIM) index

SSIM metrics uses structural distortion in video as an estimate of perceived visual distortion. It is based on an assumption that HVS is highly adapted for extraction of structural information from the viewing field. So, the level of perceived impairment is proportional to the perceived structural information loss instead of perceived errors.

For the structural distortion measure, SSIM uses means (μ and μ'), variances (σ and σ') and covariance (cov) of the original and the distorted sequences. These values are calculated for 8 x 8 pixels blocks $b_i(x,y,t)$ of the original video frame $p(x,y,t)$, and for blocks $b_i'(x,y,t)$ of the distorted video frame $p'(x,y,t)$.

The SSIM index for a block $b_i(x,y,t)$ is calculated as

$$\begin{aligned} SSIM_i(t) &= l_i(t) \cdot c_i(t) \cdot s_i(t) \\ &= \frac{4\mu_i(t) \cdot \mu_i'(t) \operatorname{cov}_i(t)}{(\mu_i^2(t) + \mu_i'^2(t)) \cdot (\sigma_i^2(t) + \sigma_i'^2(t))} \end{aligned} \quad (15)$$

where $l_i(t)$, $c_i(t)$ and $s_i(t)$ are defined as follows

$$l_i(t) = \frac{2 \cdot \mu(t)_i \cdot \mu_i'(t)}{\mu_i^2(t) + \mu_i'^2(t)} \quad (16)$$

$$c_i(t) = \frac{2 \cdot \sigma_i(t) \cdot \sigma_i'(t)}{\sigma_i^2(t) + \sigma_i'^2(t)} \quad (17)$$

$$s_i(t) = \frac{\operatorname{cov}_i(t)}{\sigma_j(t) \cdot \sigma_j'(t)} \quad (18)$$

Parameter $l_i(t)$ gives a luminance difference, $c_i(t)$ a contrast difference and $s_i(t)$ a structure difference measure between blocks of the original and the disturbed video frame.

SSIM indexes are not calculated for the entire frame, but only for properly selected R blocks, thereby reducing significantly computational cost while still providing good experimental results.

The SSIM index includes structural impairments in Y, C_b and C_r color components with different weights. The local quality index for every block is given by

$$SSIM_i(t) = 0.8 \cdot SSIM_i^Y(t) + 0.1 \cdot SSIM_i^{Cb}(t) + 0.1 \cdot SSIM_i^{Cr}(t) \quad (19)$$

Based on block $SSIM_i$, a quality index is calculated for every frame by using weighing value w_i. The authors selected w_i between 0 and 1, in dependence of local luminance $\mu_i(t)$. The frame quality index $Q(t)$ is given by

$$Q(t) = \frac{\sum_{i=1}^{R} w_i(t) \cdot SSIM_i(t)}{\sum_{i=1}^{R} w_i(t)} \quad (20)$$

Finally, the overall quality of the entire sequence is obtained as the weighted sum of frame quality indexes. Weighing value $W(t)$ for a frame at moment t depends on the motion level in that frame. For a higher level of motion in the frame model uses smaller $W(t)$ because spatial distortion is less visible in a fast moving video. Quality index *SSIM* for the entire video sequence is given by

$$SSIM = \frac{\sum_{t} W(t) \cdot Q(t)}{\sum_{t} W(t)} \quad (21)$$

Experimental results reported in [7] show that the SSIM quality metrics obtained Pearson correlation 0.830 measured on test video sequences from VQEG Phase I.

3. EXPERIMENTAL RESULTS

We have made objective and subjective measurements for two CIF video sequences, *head* and *nature*, with 29.97 frames per second, coded with an XviD coder with 7 coding rates: 16, 50, 128, 250, 750, 1,250 and 2,500 kbits/s. One frame from each original test sequence is shown in Fig. 1.

a) b)

Fig.1. Frame from the test sequence: a) *head* **;** b) *nature*

The choice of sequences is based on their very distinct contents. The *head* video presents a speaker in the central position of all frames. The *nature* video is characterized by rapid changes of content and a high level of details in every frame.

Three objective metrics are used in our experiments: PSNR, VQM and SSIM. These objective measures are obtained by using the MSU Video Measurement Tool, [9]. We have made experimental subjective quality evaluation with 12 non-experienced observers by using the MSU Perceptual Video Quality tool, [9], and the Double Stimulus Impairment Scale (DSIS) according to ITU-R BT.500-11. Results are given as average MOS for each coding rate for each sequence. Objective and subjective quality evaluation results for sequence *head* are given in Table 1. and results for sequence *nature* are given in Table 2.

Table 1. Results for sequence *head*

Bit rate (kbit/s)	PSNR	SSIM	VQM	MOS
16	29.71	0.881	1.901	1
50	30.61	0.899	1.745	1.17
128	34.76	0.959	1.177	2.17
250	36.29	0.973	0.975	3.25
750	37.70	0.986	0.749	3.75
1250	37.96	0.988	0.679	4.25
2500	38.06	0.989	0.647	4.58

Table 2. Results for sequence *nature*

Bit rate (kbit/s)	PSNR	SSIM	VQM	MOS
16	29.53	0.817	2.127	1
50	29.59	0.819	2.104	1
128	31.43	0.977	1.598	2.33
250	32.40	0.912	1.305	3.25
750	33.01	0.936	1.055	3.75
1250	33.13	0.941	0.982	4.5
2500	33.19	0.944	0.917	4.65

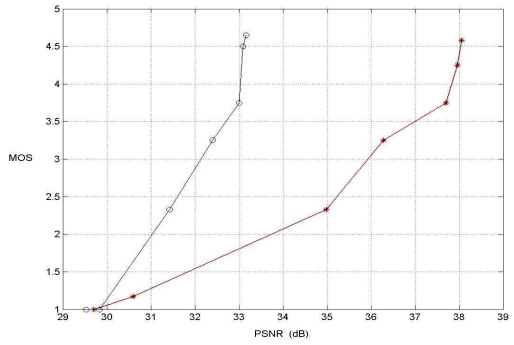
In Fig. 2. a) MOS grades versus PSNR scores are given for sequences *head* and *nature* with different bit rates. MOS grades versus SSIM scores are given in Fig. 2. b), whereas MOS grades versus VQM scores are given in Fig. 2. c).

Although the experiment is carried out with a small number of test sequences and a relatively small number of observers, some useful conclusions can be drawn.

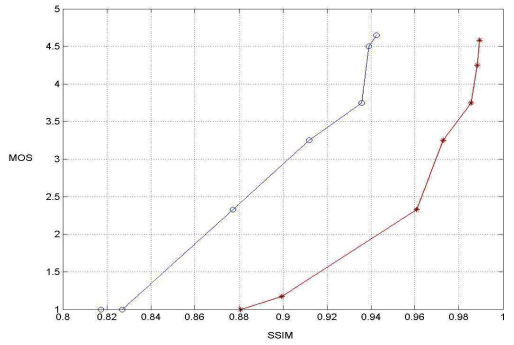
Results in Table 1. and Table 2. show that MOS grades for two sequences are very close on a given bit rate, across whole range of bit rates. Average difference between MOS grades for two sequences is 0.093, with a maximum 0.25 at 1250 kbits/s.

All objective quality metrics give significantly different results for *head* and *nature* sequences for a given bit rate, and, what is more important, for the same (or similar) MOS grade. This difference rises for higher bit rates.

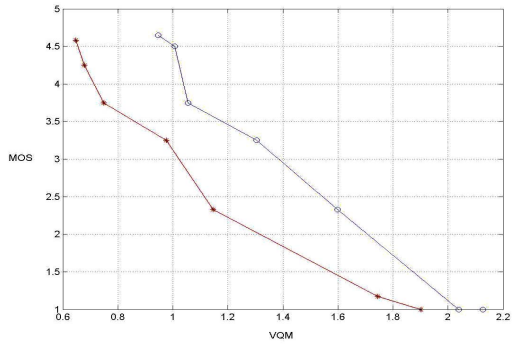
As can be seen in Fig. 2., the difference between the results for *head* and *nature* sequences are high for all three objective metrics. For the PSNR metrics the average difference is 3.26 dB, which makes 38.3% of the whole PSNR range in this experiment. For the SSIM metrics the average difference is 0.0609, i.e. 60.1% of the measured SSIM range and for the VQM metrics the average difference is 0.319, i.e. 28.1% of the measured VQM range. Although the VQM metrics gives the closest results for two videos, the difference between results is so high that one cannot say what VQM value is a threshold value



a)



b)



c)

Fig. 2. a) MOS vs. PSNR; b) MOS vs. SSIM; c) MOS vs. VQM

indicating for example the MOS score 4. The *nature* video obtains MOS score 4 for the VQM value close to 1, while the *head* video has MOS score close to 3 for the same VQM value.

Masking properties of video content, importance of content for a human observer, even the position of impairment, influence subjective experience of the distorted video, as also reported in [10].

4. CONCLUSION

Objective metrics use spatial summation over frame and temporal summation over sequence to achieve overall quality grade for a video sequence. Experimental results of objective and subjective video quality measurements for two video sequences with distinct content show that the correspondence between the objective and the subjective grades depends not only on the methods used but also on the video content. The temporal summation can be more critical because of mutual influence of spatial and temporal masking, as well as video content importance for human observer.

Objective quality measures show a high correspondence to subjective quality grades (reported correlation is close to or higher than 0.9), and can be used for evaluation of different coding techniques or different channel conditions. But for the definition of the QoS parameter needed for the required quality of the given video transmission, the spread of results is too high, and content dependence of metrics has to be improved.

REFERENCES

[1] M. Masry, S.S. Hemami, Y. Sermadevi, "A Scalable Wavelet-Based Video Distortion Metric and Applications", *IEEE Trans. on Circuits Syst. Video Technol.*, Vol. 16, No. 2, 2006, pp. 260-273

[2] A.B. Watson, J. Malo, "Video Quality Measurement Based on the Standard Spatial Observer", *Proc. ICIP*, 2002, pp. 24-28

[3] http://ise.stanford.edu/class/ee392j/projects/projects/xiao_report.pdf

[4] C.J.B. Lambrecht et al., "Quality Assessment of Motion Rendition in Video Coding", *IEEE Trans. Circuits Syst. for Video Technol.,* Vol. 9, No. 5, 1999, pp. 766-781

[5] E.P. Ong et al., "Visual Distortion Assessment With Emphasis on Spatially Transitional Regions", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 14, No. 4, 2004, pp. 559-566

[6] M.H. Pinson, S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality", *IEEE Trans. on Broadcasting*, Vol. 50, No. 3, 2004, pp. 312-322

[7] Z. Wang, L. Lu, A.C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement", *Signal Processing: Image Comm.* Vol.19, 2004, pp. 121-132

[8] Video Quality Experts Group, www.vqeg.org

[9] MSU Graphics&Media Lab, Video Group, MSU codecs, www.compression.ru/ video/

[10] M.S. Moore, J.M. Foley, S.K. Mitra, "Defect Visibility and Content Importance: Effects on Perceived Impairment", *Signal Processing: Image Communication* 19, 2004, pp. 185-203