

SUBJECTIVE EVALUATION OF A NO-REFERENCE VIDEO QUALITY MONITORING ALGORITHM FOR H.264/AVC VIDEO OVER A NOISY CHANNEL

M. Naccari, M. Tagliasacchi, S. Tubaro

Politecnico di Milano
Dipartimento di Elettronica e Informazione
20133 Milano, Italy

ABSTRACT

In this paper we evaluate NORM, a NO-Reference video quality Monitoring algorithm we proposed in a previous work, for the prediction of the subjective quality of H.264/AVC video transmitted over a noisy packet-switched network. NORM produces an estimate of the mean square error distortion at the macroblock level between the noiseless and noisy sequence, without having access to the former. The output of NORM can be readily converted into a no-reference estimate of the PSNR at the sequence level. We carried out an extensive subjective evaluation campaign on CIF and 4CIF resolution sequences encoded with H.264/AVC and transmitted over a channel that drops packets at different packet loss rates, to obtain the differential mean opinion scores. Our results show that the estimated PSNR achieves good correlation with the subjective scores, very close to the ones achieved by the PSNR computed in full-reference mode, i.e. as if the noiseless sequence would be available at the decoder.

Index Terms— Objective and Subjective video quality assessment, no-reference quality assessment.

1. INTRODUCTION

The use of IP networks for the delivery of multimedia contents is gaining an increasing popularity as a mean of broadcasting media files from a content provider to many content consumers. In the case of video, for instance, packet-switched networks are used to distribute programs in IPTV applications. Typically, these kinds of networks provide only best-effort services, i.e. there is no guarantee that the content will be delivered without errors to the final users. Therefore, the content provider might be interested in monitoring the actual perceived quality at the end-user terminal, where the original noiseless content is generally unavailable.

In practice, the received video sequence may be a degraded version of the original one. Besides the distortion introduced by lossy coding, the user's experience might be affected by channel induced distortions. In fact, the channel might drop packets, thus introducing errors that propagate along the decoded video because of the predictive nature of conventional video coding schemes [1], or it might cause jitter delay, due to decoder buffer underflows determined by network latencies. We acknowledge that both aspects are equally important in determining the perceived quality. Nevertheless we address in this paper only the effect of packet losses and we refer the reader to the available literature [2] for aspects related to the effect of delay.

In our earlier works [3][4], we proposed NORM, a NO-Reference video quality Monitoring algorithm that is able to efficiently estimate the Mean Square Error (MSE) distortion between

the noiseless and noisy video sequences, i.e. between the reconstructed sequence at the encoder and decoder side. NORM produces such an estimate at the macroblock, frame and sequence level, by parsing the received H.264/AVC bitstream to extract information about coding modes, motion vectors and prediction residuals. In [4] the estimated MSE is fed forward in a reduced-reference quality monitoring scheme that computes an approximation of the Structural SIMilarity metrics (SSIM) [5], which typically shows a good correlation with the subjective Mean Opinion Score (MOS).

Unfortunately, the bit-rate overhead imposed by the auxiliary reduced-reference channel is not negligible. Thus, a no-reference system would be preferable in many practical application scenarios. In addition, it is interesting to evaluate the correlation of objective quality metrics directly with MOS, instead of using SSIM or other metrics as proxies. Motivated by these objectives, the main contribution of this paper is two-fold. First, we collected MOS values in a formal subjective evaluation campaign. A group of subjects rated the perceived quality of video contents transmitted over a noisy channel characterized by a Packet Loss Rate (PLR) in the interval [0.1% – 10%]. Second, we compared two no-reference metrics, i.e. NORM and W-NORM, which represent no-reference approximations of PSNR and SSIM, together with their full-reference counterparts.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the state-of-the-art on objective video quality monitoring. Section 3 introduces subjective quality assessment and illustrates the test material, environmental setup and subjective evaluation process. Section 4 briefly summarizes NORM and illustrates how to compute the W-NORM metrics. Section 5 discusses the analysis of the correlation between the collected subjective scores and various no-reference and full-reference metrics. Concluding remarks are given in Section 6.

2. RELATED WORK

The quality of a video sequence can be evaluated with algorithms that compute metrics that aim to predict the MOS value on the basis of some features extracted from decoded videos [6]. Objective quality metrics can be classified in: full-reference, reduced-reference and no-reference, based on the amount of information available for comparison with the original content. Full-reference metrics require the entire reference content to be available at the receiver side. Reduced-reference metrics rely on a coarse, feature-based representation of the transmitted video which is available at the receiver side without errors. No-reference objective metrics estimate the received video quality without any information on the error-free video content. Therefore, these metrics must rely on some assumptions as, for

example, the type of distortion introduced in transmitted videos.

The most commonly used full-reference objective metrics for video is the peak signal-to-noise ratio (PSNR), which is based on the mean square error between the original and the received frames. Despite its widespread diffusion as distortion metric, due to its simplicity of calculation and the easiness to deal with in optimization, it is known that in some cases, such as degradations introduced by lossy coding, PSNR could be poorly correlated with the actual perceived quality. Therefore, there is a strong interest in designing new objective quality metrics that show higher correlation with the human perceived video quality. Some of these techniques have been validated by the Video Quality Experts Group (VQEG) [7], which has tested several video quality measurement systems in the VQEG Full Reference Television (FRTV) phase 2 tests. More recently, Wang et al. have proposed a Structural SIMilarity index (SSIM) for evaluating the perceptual quality of degraded images [5]. SSIM tries to quantify the distortion in terms of the perceived loss of “structure” in the image. In [8] these ideas have been extended to the case of video sequences to generate a metric that we denote as Video SSIM (VSSIM) in this paper. The results reported by the authors show a very high correlation with MOS and a lower outlier ratio than competing methods, which make this metric one of the most promising perceptual evaluation criteria for video quality assessment.

Video quality assessment is more complex when the original reference is unavailable, as in no-reference systems. The work in [9] describes an algorithm to estimate the MSE distortion for bitstreams coded with any conventional motion-compensated video codec by considering temporal error propagation. The system proposed in [10] derives a model to estimate the received video quality taking into account parameters such as the used codec, the adopted error concealment strategy, the bit-rate and the packetization used. The work in [11], proposes a no-reference MSE estimation algorithm that measures the error concealment effectiveness on the basis of motion information and boundary distortion between the lost macroblock and its neighbors. A different approach is pursued in [12], where machine learning classifiers are used to predict packet loss visibility in MPEG-2 coded bitstreams.

3. SUBJECTIVE VIDEO QUALITY ASSESSMENT

In subjective tests, a group of people is asked to watch a set of video clips and rate their quality. The scores assigned by the observers are averaged in order to obtain the mean opinion score (MOS). In order to produce meaningful MOS values, the test material needs to be carefully selected and the subjective evaluation procedure must be rigorously defined. In our work, we adapted the specifications given in [13].

3.1. Test video sequences

In our subjective evaluation campaign we adopted six video sequences. All the original sequences are available in progressive format at a frame rate of 30 fps. Three sequences are at CIF spatial resolution (352×288 pixels), namely *Foreman*, *Coastguard* and *News*, and three at 4CIF resolution (704×576 pixels), namely *Harbour*, *City* and *Crew*. These sequences have been selected since they are representative of different types of motion content and spatial complexity.

To encode the original bitstreams, we adopted the H.264/AVC reference software, version JM14.2 [14]. We encoded all sequences using the H.264/AVC main profile to enable B-pictures and CABAC for improved coding efficiency. The GOP size has been set equal

to 15, with two B-pictures between two I or P-pictures. We did not enable specific error resiliency tools, since the goal of this work is not to evaluate the subjective impact of such tools, rather to produce noisy bitstreams and sequences from which objective quality metrics can be computed and validated against MOS. Each frame is divided into a fixed number of slices, where each slice consists of a full row of macroblocks. Rate control has been disabled since the algorithm embedded in the reference software introduced visible quality fluctuations along time for some of the tested video sequences. Instead, a fixed quantization parameter QP has been carefully selected for each sequence as to ensure good visual quality in the absence of packet losses. The average encoding rate turned out to be in the range 400-550 kbps for CIF sequences and 1700-2000 kbps for 4CIF sequences. The study of the subjective impact of the distortion due to both compression and packet losses is left to future investigations.

For each of the six original H.264/AVC bitstreams corresponding to the test sequences, we generated a number of corrupted bitstreams, by dropping packets according to a given error pattern. To simulate burst errors, the patterns have been generated at six different PLR [0.1%, 0.5%, 1%, 3%, 5%, 10%] with a two state Gilbert’s model [15]. We tuned the model parameters to obtain an average burst length of 3 packets, since it is characteristic of IP networks [16]. We manually selected four channel realizations for each PLR, for a total of 24 realizations per video sequence, in order to uniformly span a wide range of distortions (and perceived video quality levels), measured in terms of PSNR. Each bitstream is decoded with the H.264/AVC reference software decoder with motion-compensated error concealment turned on, according to the implementation included in the H.264/AVC reference software [17].

3.2. Subjective evaluation procedure

Each test session involves only one subject per display assessing the test material. Subjects are seated directly in line with the center of the video display at a specified viewing distance, which is equal to $6-8H$ ($4-6H$) for CIF (4CIF) resolution sequences, where H is the height of the video window. The test environment has been acoustically isolated and the display properly calibrated.

In our subjective evaluation we adopt a single-stimulus method in which a processed video sequence is presented alone. In addition, we applied the “hidden reference removal” procedure, so that the score corresponding to the noiseless sequence can be subtracted during data analysis to obtain Differential Mean Opinion Scores (DMOS).

At the end of each test sequence, human subjects provide a quality rating using a continuous rating scale in the range [1 – 5]. Each subjective experiment refers to a single spatial resolution and includes the same number of 80 sequences: 3×24 realizations of each of the three test sequences, 3 hidden references and 5 practice clips, shown at the beginning of the experiment to allow the viewer to familiarize with the assessment procedure. The presentation order is randomized according to a random number generator, discarding those permutations where there are more than three consecutive realizations of the same sequence [13].

3.3. Subjective data analysis

Each video content is analyzed independently. For each of the S subjects, we store the difference between the score assigned to the hidden reference and to each of the 24 realizations, obtaining a $24 \times S$ matrix of differential raw scores. The ANOVA statistical model is applied to verify whether there is a significant disagreement between

the mean scores assigned by the different individuals. In this case normalization is performed as described in [18]. Then, the normalized raw scores are tested for the presence of outliers, i.e. subjects that tend to assign scores that are significantly different from the average population. Finally, DMOS scores are computed for each of the 24 realizations by computing the average across the remaining subjects, which in our case are always more than 15.

4. NORM AND W-NORM

The NORM algorithm aims at estimating the channel induced distortion at the macroblock granularity according to the mean square error (MSE) metrics:

$$D_n^i = \frac{1}{B^2} \sum_{p=1}^B \sum_{q=1}^B \left(\mathbf{X}_n^i(p, q) - \mathbf{Y}_n^i(p, q) \right)^2, \quad (1)$$

where \mathbf{X}_n^i denotes the i -th $B \times B$ macroblock in frame n reconstructed at the encoder side (i.e. available at the decoder in the error free scenario), and \mathbf{Y}_n^i the same macroblock reconstructed at the decoder side when channel losses occurred. Since the MSE is an additive metric, it can be readily computed at frame or sequence level by summing up the contributions of the individual macroblocks.

The channel induced distortion is modeled distinguishing between whether a macroblock has been correctly received or not and taking into account the predictive (both spatial and temporal) nature of the H.264/AVC codec, together with the concealment algorithm adopted by the decoder. In fact, when a macroblock is correctly received, the decoder can reconstruct its pixel values, although errors might propagate from frames used as reference in the motion-compensation phase (temporal error propagation) or from neighboring reconstructed pixel values in the same frame (spatial error propagation). Conversely, when a macroblock is lost, errors are inevitably introduced in the areas corresponding to the missing pixels. Furthermore, depending on the decoder concealment strategy, errors can propagate either along the temporal or spatial dimension. Therefore, we can envisage the following scenarios: a) Correctly received intra predicted macroblock; b) Correctly received inter predicted macroblock; c) Lost macroblock and spatial concealment; d) Lost macroblock and temporal concealment. In [3] we provided a detailed explanation on how to estimate the channel-induced distortion (1) for the four cases above. The NORM score is computed at sequence level by averaging the estimated MSE distortion at macroblock level and converting in a log-scale:

$$\text{NORM}(\mathbf{X}, \mathbf{Y}) = 10 \log_{10} \frac{255^2}{\frac{1}{NL} \sum_{n=1}^N \sum_{i=1}^L \hat{D}_n^i}, \quad (2)$$

where N and L are, respectively, the number of video frames and the macroblocks for each frame. Conversely, the proposed W-NORM index is computed by averaging the following scores obtained at the macroblock level

$$\text{W-NORM}(\mathbf{X}_n^i, \mathbf{Y}_n^i) = \min \left(1, \max \left(0, 1 - \frac{\hat{D}_n^i}{2\sigma_{\mathbf{Y}_n^i}^2 + C_2} \right) \right), \quad (3)$$

where C_2 is a normalization constant (we set $C_2 = 0.03$). Therefore the MSE distortion is weighted by the local variance of each macroblock. This normalization is intuitively justified by observing that artifacts introduced by packet losses are more visible in areas characterized by smooth textures.

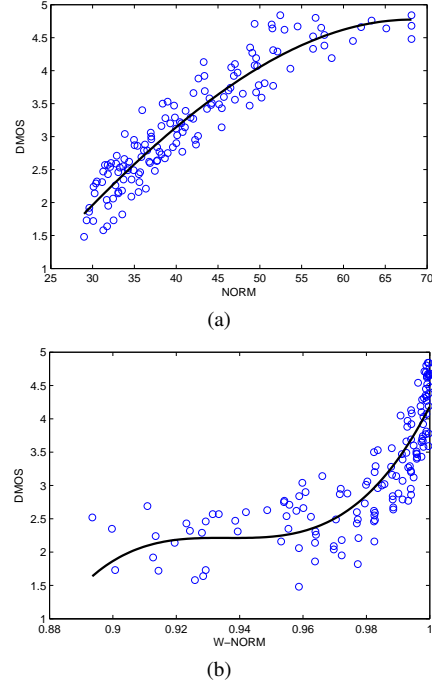


Fig. 1. a) DMOS vs. NORM. b) DMOS vs. W-NORM.

It can be easily shown that W-NORM is equivalent to a simplified version of the SSIM index, computed in no-reference mode. From [5], the SSIM index computed between two macroblocks \mathbf{X} and \mathbf{Y} is defined as

$$\text{SSIM}(\mathbf{X}_n^i, \mathbf{Y}_n^i) = \frac{\left(2\mu_{\mathbf{X}_n^i} \mu_{\mathbf{Y}_n^i} + C_1 \right) \cdot \left(2\sigma_{\mathbf{X}_n^i, \mathbf{Y}_n^i} + C_2 \right)}{\left(\mu_{\mathbf{X}_n^i}^2 + \mu_{\mathbf{Y}_n^i}^2 + C_1 \right) \cdot \left(\sigma_{\mathbf{X}_n^i}^2 + \sigma_{\mathbf{Y}_n^i}^2 + C_2 \right)}, \quad (4)$$

Equation (3) can be obtained from (4) by neglecting the first factor (i.e. assuming that $\mu_{\mathbf{X}_n^i} = \mu_{\mathbf{Y}_n^i}$). In fact, with little algebra we obtain an expression of the sample covariance by means of the MSE D_n^i

$$\sigma_{\mathbf{X}_n^i, \mathbf{Y}_n^i} = \frac{\left(\mu_{\mathbf{X}_n^i}^2 + \mu_{\mathbf{Y}_n^i}^2 + \sigma_{\mathbf{X}_n^i}^2 + \sigma_{\mathbf{Y}_n^i}^2 \right) - \left(D_n^i + 2\mu_{\mathbf{X}_n^i} \mu_{\mathbf{Y}_n^i} \right)}{2}. \quad (5)$$

By further assuming $\sigma_{\mathbf{X}_n^i}^2 = \sigma_{\mathbf{Y}_n^i}^2$ (since $\sigma_{\mathbf{X}_n^i}^2$ is unavailable in no-reference mode), (4) reduces to (6). The W-NORM score is computed at sequence level by averaging the estimated W-NORM at the macroblock level:

$$\text{W-NORM}(\mathbf{X}, \mathbf{Y}) = \frac{1}{NL} \sum_{n=1}^N \sum_{i=1}^L \text{W-NORM}(\mathbf{X}_n^i, \mathbf{Y}_n^i). \quad (6)$$

5. RESULTS AND DISCUSSION

We compared four metrics. Two no-reference metrics defined in the previous section, namely NORM (2) and W-NORM (6), and two full-reference metrics, namely PSNR and SSIM [5]. In the latter cases, the reference is assumed to be the noiseless reconstructed sequence, rather than the original uncoded one.

Table 1. Performance indicators of the tested metrics.

	ρ	σ	RMSE
PSNR	0.945	0.944	0.286
NORM	0.938	0.939	0.301
SSIM	0.912	0.935	0.357
W-NORM	0.868	0.889	0.432

In order to evaluate the prediction capability of the tested metrics, a non-linear mapping step is applied before computing any performance indicator. As in [13], we adopted a cubic polynomial

$$\widehat{\text{DMOS}}(x) = ax^3 + bx^2 + cx + d, \quad (7)$$

where $\widehat{\text{DMOS}}(x)$ represents the predicted DMOS value and x is the value assumed by any tested metrics (NORM, W-NORM, PSNR or SSIM). The optimal polynomial parameters for each metrics are obtained by minimizing the mean square error between the observed DMOS and the predicted $\widehat{\text{DMOS}}$ values, while guaranteeing the fitting to be monotonic. Figure 1 illustrates the fitted polynomials for the case of NORM and W-NORM, respectively.

As performance indicators we computed both the Pearson's (ρ) and Spearman's (σ) correlation coefficients. The former measures the linear relationship between a model's performance and the subjective data, i.e. between the measured and predicted DMOS. The latter indicates the correlation between the ranking order of DMOS and $\widehat{\text{DMOS}}$. Both indicators assume value in the $[-1, 1]$ range, where a value close to 1 indicates a higher (positive) correlation. We also included in the analysis the RMSE (Root Mean Square Error) metrics, which computes the square root of the average error $\text{DMOS} - \widehat{\text{DMOS}}$.

Table 1 summarizes the three performance indicators for the four tested metrics when DMOS scores attributed to all tested sequences are aggregated. Due to space limitations we cannot present the results obtained by analyzing independently CIF and 4CIF sequences.¹ We notice that, as expected, the full-reference metrics performs better than the corresponding no-reference metrics. In fact, PSNR achieves a value of $\rho = 0.945$, slightly higher than NORM $\rho = 0.938$, while SSIM obtains $\rho = 0.912$ vs. $\rho = 0.868$ of W-NORM. These preliminary results suggest the following observations: a) For the problem at hand, the weighting scheme adopted by W-NORM does not seem to be beneficial. In fact, W-NORM is inspired by SSIM, which has been originally developed and tested on degradations affecting the frames somewhat uniformly in space, e.g. compression artifacts. Conversely, packet losses tend to be localized in space and their visibility depends probably more on the spatio-temporal extent of the loss than on the amount of local image structure. b) NORM achieves almost the same performance as the full-reference PSNR, while working in no-reference mode. This was predictable, since we already showed in our earlier work [3][4] that NORM estimates the true mean square error distortion very accurately, especially at the sequence level. Overall, we can claim that NORM can be effectively used to accurately predict the visual quality of H.264/AVC sequences transmitted over an error prone channel.

6. CONCLUSIONS

This paper shows that the estimated means square error distortion provided by NORM can be effectively used to predict the visual

¹Supplementary results are available at the author's web page: <http://home.dei.polimi.it/tagliasa/suppl.icip2009.pdf>

quality of sequences transmitted over a noisy channel. Future work will investigate novel weighting schemes to be applied to the macroblock level estimate of NORM to consider the visual impact spatio-temporally correlated losses aimed at further improving the prediction performance.

7. REFERENCES

- [1] I. E. G. Richardson, *Video Codec Design*, John Wiley & Sons, 2002.
- [2] M. Claypool and J. Tanner, "The effects of jitter on the perceptual quality of video," in *ACM Multimedia*, Orlando, FL, USA, November 1999.
- [3] M. Naccari, M. Tagliasacchi, F. Pereira, and S. Tubaro, "No-reference modeling of the channel induced distortion at the decoder for H.264/AVC video coding," in *Proceedings of the International Conference on Image Processing*, San Diego, CA, USA, October 2008.
- [4] A. Albonico, G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro, "A reduced-reference video structural similarity metric based on no-reference estimation of channel-induced distortion," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Taipei, TW, April 2009.
- [5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.
- [6] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Trans. Broadcast. (to appear)*, 2008.
- [7] "The Video Quality Expert Group web site," <http://www.its.bldrdoc.gov/vqeg>.
- [8] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measure," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, February 2004.
- [9] A. R. Reibman, V. A. Vaishmpayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 327–334, April 2004.
- [10] T. Shu, J. Apostolopoulos, and R. Guérin, "Real-time monitoring of video quality in IP networks," in *International Workshop on Network and Operating System Support for Digital Audio and Video*, Stevenson, WA, USA, June 2005.
- [11] T. Yamada, Y. Miyamoto, and M. Serizawa, "No-reference video quality estimation based on error-concealment effectiveness," in *IEEE Packet Video*, Lausanne, Switzerland, November 2007.
- [12] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishmpayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 341–355, April 2006.
- [13] Video quality expert group, "Final report from the Video Quality Expert Group on the validation of objective models of multimedia quality assessment, phase 1," Tech. Rep., VQEG, September 2008, Version 2.6.
- [14] Joint Video Team (JVT), "H.264/AVC reference software version JM14.2," downloadable at <http://iphome.hhi.de/suehring/tml/download/>.
- [15] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell System Technical Journal*, vol. 39, pp. 1253–1266, September 1960.
- [16] T.-K. Chua and D. C. Pheanis, "QoS evaluation of sender-based loss-recovery techniques for VoIP," *IEEE Netw.*, vol. 20, no. 6, pp. 14–22, December 2006.
- [17] G. J. Sullivan, T. Wiegand, and K.-P. Lim, "Joint model reference encoding methods and decoding concealment methods," Tech. Rep. JVT-I049, Joint Video Team (JVT), September 2003.
- [18] E. Drelie Gelasca, "Full-reference objective quality metrics for video watermarking, video segmentation and 3D model watermarking," Tech. Rep., EPFL, September 2005, Ph.D. Thesis.