5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)
29-31 August 2016, Berlin, Germany

Comparison of EWPSNR and MOS on an Eye-tracking Labelled Video Dataset

Saman Zadtootaghaj [1], Hamed Ahmadi [2], Sebastian Möller [3]

[1] Telekom Innovation Labs, Deutsche Telekom AG, Germany
[2] Multimedia Processing Laboratory (MPL), University of Tehran, Tehran, Iran
[3] Quality and Usability Lab, TU Berlin, Germany

saman.zadtootaghaj@telekom.de, ha.ahmadi@ut.ac.ir, sebastian.moeller@telekom.de

Abstract

Perceptual video compression has been increasingly utilized to obtain higher compression gains by removing perceptual redundancies according to the Human Visual System (HVS). Due to non-linear complex mechanisms of the HVS, conventional quality metrics fail to assess the performance of perceptual video compressors. Therefore, subjective assessment is employed as the most reliable method for this purpose. However, it costs both time and money which consequently impedes perceptual compression development. There have been several attempts to develop perceptual objective metrics as an alternative to subjective assessments. Among them, Eye-tracking Weighted PSNR (EWPSNR) is believed to be the fairest where the gaze points are available. In this paper, we measure the correlation between EWPSNR and Mean Opinion Score (MOS) on an eye-tracking dataset to evaluate the performance of EWPSNR to predict the subjective quality.

1. Introduction

With the support of advances in related technologies, high data-size video-based applications have been drastically utilized in various domains such as entertainment, sports and education. Modern video coding standards condense this increasing visual data as much as possible while minimizing the loss of visual quality due to compression. They employ multifarious methods to remove both statistical and perceptual redundancy. However, since the perceptual properties of the human visual system (HVS) have not yet been fully understood, it may still be possible to massively improve the coding efficiency without provoking a significant perceptual quality degradation [1].

The main idea of almost all perceptual coders is to maximize perceived quality. In order to do so, estimations of quality which reflect true perception rather than a conventional peak signal-to-noise ratio (PSNR) or the mean square error (MSE) are necessary. The basic difference between conventional and perceptual quality measures is that the former looks at distortions from the signal point of view, which leads to poor correlation with perceived quality [2], whereas the latter takes the perceptual point of view. Therefore, the efficacy of perceptual quality metrics is the key to evolution of more efficient perception-aware coding techniques.

In this paper, we first review the existing perceptual quality metrics and then evaluate eye-tracking-weighted PSNR (EWPSNR) [3] as the fairest measure when the gaze points are available, because it has its roots in the psychophysical properties of the HVS rather than the coding [1]. To do so, we measure the Pearson correlation coefficient between EWPSNR and Mean Opinion Score (MOS) for an eye-tracking dataset. In addition to the reference video sequences, this dataset also contains the eye-tracking data and MOS for video sequences distorted in different ways. Therefore, it would be possible to assess EWPSNR under varied circumstances.

The results of this paper not only are interesting for designing a new perceptual quality metric, but also can be valuable for measuring the impact of user behavioral information, namely eye movements, as a quality of experience (QoE) factor. Therefore, it could help to propose novel video coders which achieve low bit rate and high quality of experience in the same time.

This paper is organized as follows. In the next section, perceptual video metrics are reviewed. Section 3 introduces the eye-tracking dataset that we utilized in our experiments. In Section 4, the experiments' details and results are explained. Section 5 discusses the results and Section 6 concludes the paper.

2. Related work

Most perceptual objective metrics are designed based on the basic form of PSNR, while weighted versions of MSE are used in order to consider perceptual properties in the weights [3, 4, 5, 6, 7, 8]. Examples include the foveal PSNR (FPSNR) [4], which weighs distortions with the local bandwidth decreasing with eccentricity, and peak signal-to-perceptible noise ratio (PSPNR) [6] and foveated peak signal-to-perceptible noise ratio [7], which consider only errors greater than JND and foveated JND thresholds, respectively. Among these metrics, EWPSNR is considered a fair way because it is calculated directly based on eye-tracking data and does not depend on any perceptual models [1]. The corresponding computation formulas are as follows.

$$EWMSE = \frac{1}{\sum_{x=1}^{M}\sum_{y=1}^{N} w_{x,y}} \sum_{x=1}^{M}\sum_{y=1}^{N} (w_{x,y} \cdot (I'_{x,y} - I_{x,y})^2) \quad (1)$$

$$EWPSNR = 10 * \log\left(\frac{(2^n-1)^2}{EWMSE}\right) \quad (2)$$

$$w_{x,y} = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\left(\frac{(x-x_e)^2}{2\sigma_x^2} + \frac{(y-y_e)^2}{2\sigma_y^2}\right)} \quad (3)$$

where I and I′ are the original frame and the encoded frame, respectively, M and N are the frame's height and width in pixels, n is the bit depth of the color component, and $w_{x,y}$ is the weight for distortion at position (x, y). $w_{x,y}$ is computed based on the subjects' eye fixation position (x_e , y_e)

10.21437/PQS.2016-20

from eye-tracking experiment, σ_x and σ_y are two parameters related to the distance and view angle, usually taken from fovea size. EWPSNR uses a Gaussian distribution around each eye fixation point for weighting. This kind of metric is biologically inspired by the fovea mechanism. The fovea is a circular region of about 1.5 mm in diameter on the retina in which the density of sensor cells is the highest and decreases rapidly with respect to the angle with the visual axis, called eccentricity.

Several research papers analyzed the relation between conventional full-reference objective metrics and subjective assessment [9, 10, 11]. In [10], the performance of four video quality metrics (SSIM, MS-SSIM, VQM, and MOVIE), together with their modified versions, which had saliency maps incorporated to their algorithms, have been compared. Our work differs from theirs, in the sense that we have directly used eye-tracking data to determine perceptually important areas rather than using saliency models.

Usually the individual value of a video quality metric for each frame is averaged over a complete video sequence, producing one value representing the visual quality of the whole sequence. Temporal pooling is though utilized to improve the performance of the visual quality metrics [12]. In this paper, we show that the same approach should be taken into account for EWPSNR.

3. Dataset

For our analysis, we selected a dataset which was presented in [12]. This dataset includes subjective quality assessment data. The dataset contains eye tracking data which was gathered while showing 20 standard videos (720*576, interlaced, 50 Hz) with five levels of quality. To generate the test sequences, quality of each reference video was coded with the H.264/AVC codec, and then a simulated transmission with four transmission errors in varied spatial position and duration were applied. Two values of transmission errors were applied to Regions Of Interest (ROI) and two other errors were applied in non-ROI regions. For each video content, the reference video sequence along with four reduced-quality videos were shown to participants who had to rate the visual quality on a 5-point impairment scale. The average Mean Opinion Scores (MOS) for each video sequence was collected from 30 non-expert participants (10 female, 20 male) with an average age of about 23 years. The experiment was designed according to ITU Rec. BT.500 [14]. The videos were presented on a LVM-401W full HD. screen by TVlogic with a size of 40" and a native resolution of 1920 × 1080 pixels. A mid-grey background was added to the SD test sequences to be displayed on the HD screen. The observers were seated at a distance of about 150 cm corresponding to six times the height of the used display area. The 5-point impairment scale [14] was used to assess the annoyance of the distortions in the sequences. Here, the observers assigned one of the following adjectival ratings to each of the sequences: 'Imperceptible (5)', 'Perceptible, but not annoying (4)', 'Slightly annoying (3)', 'Annoying (2)', and 'Very annoying (1)'.

4. Experiment

The main goal of our experiments is to analyze how accurate EWPSNR, as a perceptual objective metric, can predict subjective quality. Indeed, it seems that such a perceptual video metric should have a better correlation with perceived quality in comparison with other typical metrics like PSNR, since EWPSNR makes use of additional user information, namely eye movements. As a measure of the subjective quality, we use MOS values. We first calculate EWPSNR for each video sequence in the dataset and then measure the correlation among the calculated EWPSNRs and MOS values which are already included in the dataset. Before doing so, we need to find a suitable way to calculate EWPSNR of a whole video sequence, because EWPSNR was originally proposed to calculate the perceptual quality on a frame by frame basis and thus, does not consider the video quality variations over time. Figure 1 depicts the EWPSNR variations for one subject viewing three different versions of an exemplary video sequence in the dataset. Quality variations exist even among the frames of the reference video sequence, although with less amplitudes. A straightforward method is to calculate the average EWPSNR over the frames. However, our experimental results let us assume that two other factors, namely temporal development of quality and the effect of the worst quality frame, have their own impact and should also be covered in the final formula. A similar work on PSNR presented the shortcomings of not considering the variations of the video quality metric over time (frames) [13].

In Figure 2, the correlation between the EWPSNR values of the i^{th} frames of all the video sequences and the MOS values were calculated. As can be seen, a video sequence's frames do not equally contribute to the user perceived quality. The results show that two sets of frames have a huge impact on the user perceived quality. The location of these influential frames differs from video to video. In our dataset, two kinds of transmission error were applied to two different spatial regions

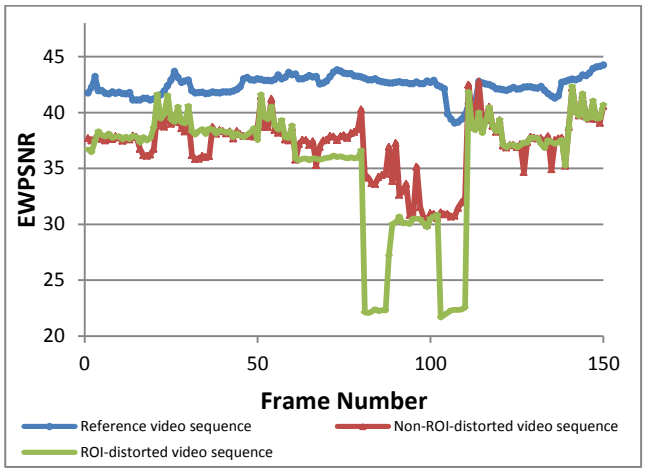

Figure 1: *EWPSNR variations over time for a selected subject on a selected video sequence.*

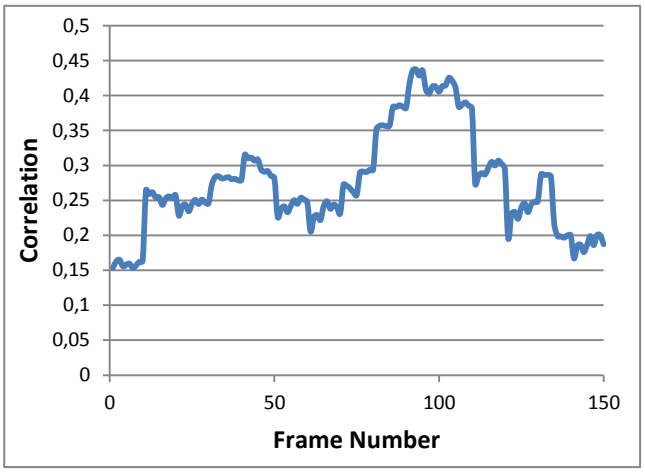

Figure 2: *Impact of each frame's EWPSNR on the MOS.*

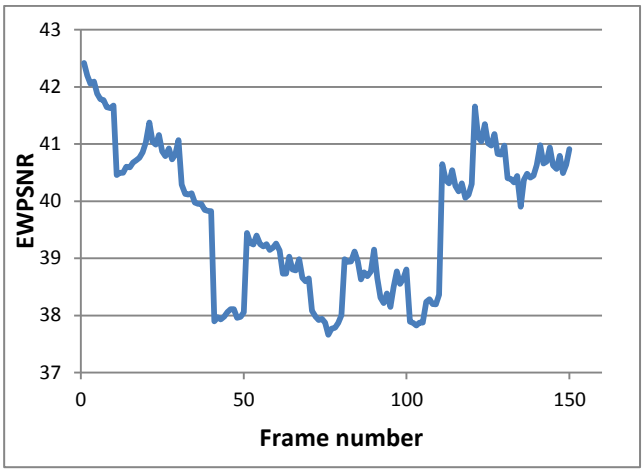

Figure 3: *Temporal distribution of the distortions*

of each video (ROI and non-ROI). Therefore, there are four different scenarios. To illustrate the temporal distribution of the distortions, we calculated the average amount of distortion in each frame number (as shown in Figure 3). To do so, for each frame number, we averaged the EWPSNR values over all subjects and video sequences.

Comparing Figure 2 and Figure 3 highlights the fact that frames with relatively more distortions have stronger impacts on user perceived quality. In the other words, the value of EWPSNR in the low-quality frames is more correlated to the MOS than that of in the high-quality frames. Therefore, users tend to judge the quality of a video sequence based on the worst quality frames they experience in that sequence.

Based on these findings, we utilized two different functions to calculate the EWPSNR for the whole video: minimum and weighted average. For the former, we simply found the minimum value EWPSNR among all frames as shown in Equation 4, where VQ_{seq} is the video quality representation of the whole video sequence, $EWPSNR_{f_i}$ is the EWPSNR value of the i^{th} frame, and N is the total number of frames in the video sequence. As for the latter, we first normalized the values in Figure 2 and then applied them as weight coefficients in the process of averaging EWPSNR values as shown in Equation 5, where w_i is the correspondent weight for the i th frame. There are several other functions proposed in the literature, such as in [12], but since their performance was not competitive we didn't report them here.

$$VQ_{seq} = \min\left(EWPSNR_{f_1}, \ldots, EWPSNR_{f_N}\right) \quad (4)$$

$$VQ_{seq} = \frac{\sum_{i=1}^{N} w_i \times EWPSNR_{f_1}}{\sum_{i=1}^{N} w_i} \quad (5)$$

Table 1 compares the correlation of these functions as well as standard average with the MOS values. As can be seen, these two functions outperform the standard average function in terms of how much they are close to the MOS values. The minimum function has the best performance. One reason to this observation is that the amount of distortion in the worst frames, in the utilized dataset, is so much that it remains in the user's mind and affects his/her judgment.

Another interesting observation, implied by comparing Figure 2 and Figure 3, is that although the amount of distortion for the two sets of influential frames is almost the same, the impact of the set at the end of the video sequences is more than that of the set at the beginning of the video sequences. To further investigate the matter, we divided the video sequences into two halves and calculated the correlation of each part on the users' opinion scores. Figure 4 shows the average EWPSNR of each half over all videos for each subject. As can be seen, the video quality of the second half of the videos plays a more significant role on user's opinion scores in the utilized dataset than that of the first half. It can be characterized as retrospective appraisal, i.e. remembered experience, or more generally the temporal development of the quality [14]. This observed effect is often called recency effect, i.e. quality events closer to the end of an episode have a stronger impact on the episode-final rating. This effect has not yet been fully understood in the field of video quality assessment. For example, while the beginning and ending frames are considered to significantly affect the user perceived quality in [13], the findings in [14] do not support the idea. Therefore, more research is required to validate the existence of the recency effect and find an efficient method to integrate it into the video quality metrics.

As mentioned before, EWPSNR was built on the grounds of discovered facts about the fovea. The size of fovea differs among people ranging from 2° to 5° of visual angle around the center of gaze. As part of our experiment, we investigate the impact of this factor on the accuracy of EWPSNR. To do so, we first divide the video sequences into three groups: reference videos, ROI-distorted videos, and non-ROI-distorted videos. Then, we calculate the average value of Equation 4 over each group's video sequences. We repeat this process for

Table 1. *The correlation between different functions of EWPSNR and MOS.*

Function	Correlation	R^2
Standard Average	42.95	1,34E-04
Weighted Average	56.48	1,12E-6
Minimum	75.16	3.41E-13

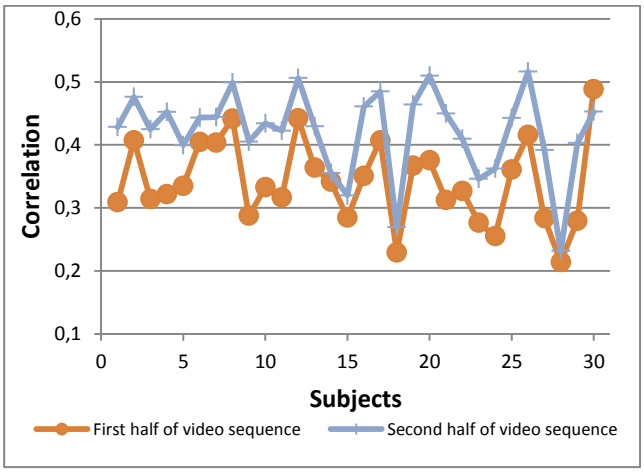

Figure 4: *Impact of video quality in each half of the video sequences on users' opinion scores.*

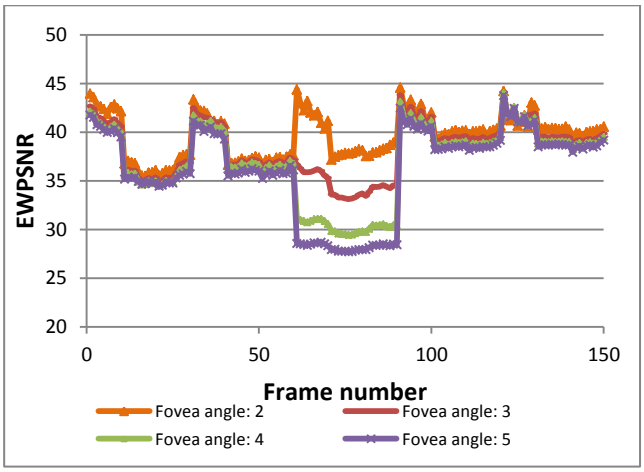

Figure 5: *EWPSNR of a selected subject and video for different fovea angles*

different values of fovea angles. Figure 5 shows EWPSNR values of an exemplary video's frames for a selected subject. The video belongs to the non-ROI-distorted group. According to the figure, different fovea angles have not significantly changed EWPSNR values in non-distorted regions of the video. However, EWPSNR values have experienced significant variations in distorted regions (approximately from frame number 60 to 90). The cause of this observation is that a larger fovea angle captures a wider portion of the screen and assigns more weights to it. Since distortions most likely exist in this wide portion, they adversely decrease the EWPSNR values. However, the final quality of the video sequences depends on the location of the distortions and gazes. Table 2 shows the correlation between EWPSNR and MOS for different fovea angles. The standard deviation of these correlation values is about 0.02 which is not significant. Therefore, measuring the participating subjects' fovea angle does not seem to be necessary.

Table 2. *The correlation between EWPSNR and MOS for different fovea angles.*

Fovea angles (°C)	Correlation	R^2
2	0.74	1.17E-13
3	0.72	9.79E-13
4	0.75	3.41E-13
5	0.69	3.78E-11

5. Discussion

The preliminary results of our experiments show that although EWPSNR is capable of predicting MOS to an acceptable extent, to become a highly accurate perceptual objective model, it still requires further improvements. This gap can be partially characterized by the fact that the subjects were reported their opinion score values in integer numbers.

In addition, we hypothesize that incorporating contextual factors, both user and system factors, into EWPSNR would increase its performance. More specifically, since the dataset was prepared under laboratory circumstances, the accuracy of participants' judgment could be influenced by the lack of several other factors. For example, the duration of all video sequences is only six seconds, which is seldom enough for the user to get immersed. We consider immersion to be very important especially in highly interactive applications such as gaming, because immersion makes users more focused and hence, restricts their attention patterns. User's awareness of the eye-tracking experiment's goal may have an adverse impact on user judgment as well.

6. Conclusions

In this paper, we showed that how reliable EWPSNR is to predict subjective quality in overall as well as how it can be improved by including information on the fovea size, and on the temporal development of quality. Further research should be conducted on eye-tracking datasets of various other applications, such as cloud gaming and video conferencing, in order to determine whether the EWPSNR accuracy is application-dependent or not. The results of such experiments pave the way to develop more profound perceptual metrics which may ultimately lead to more efficient perceptual video coders.

7. Acknowledgment

This work was funded from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 643072, Network QoE-Net.

8. References

[1] J.-S. Lee and T. Ebrahimi, "Perceptual video compression: A survey," *Selected Topics in Signal Processing,* vol. 6, no. 6, pp. 684-697, 2012.

[2] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Transactions on Broadcasting,* vol. 54, no. 3, pp. 660-668, 2008.

[3] Z. Li, S. Qin and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing,* vol. 29, no. 1, pp. 1-14, 2011.

[4] S. Lee, M. S. Pattichis and A. C. Bovik, "Foveated video quality assessment," *IEEE Transactions on Multimedia,* vol. 4, no. 1, pp. 129-132, 2002.

[5] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Transactions on Image Processing,* vol. 10, no. 10, pp. 1397-1410, 2001.

[6] C.-H. Chou and C.-W. I. Chen, "A perceptually optimized 3-D subband codec for video communication over wireless channels," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 6, no. 2, pp. 143-156, 1996.

[7] Z. Chen and C. Guillemot, "Perceptually-friendly H. 264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, no. 6, pp. 806-819, 2010.

[8] A. Cavallaro, O. Steiger and T. Ebrahim, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 15, no. 10 , pp. 1200-1209, 2005.

[9] M. C. Farias and W. Y. Akamine, "On performance of image quality metrics enhanced with visual attention computational models," *Electronics letters ,* vol. 48, no. 11 , pp. 631-633, 2012.

[10] W. Y. Akamine and M. C. Farias, "The added value of visual attention in objective video quality metrics," *Video Processing and Quality Metrics (VPQM),* 2014.

[11] W. Y. Akamine and M. C. Farias, "Video quality assessment using visual attention computational models," *Journal of Electronic Imaging,* vol. 23, no. 6 , pp. 061107-061107, 2014.

[12] C. Keimel and K. Diepold, "Improving the prediction accuracy of PSNR by simple temporal pooling," in *Video Processing and Quality Metrics for Consumer Electronics-VPQM*, 2010.

[13] U. Engelke, M. Barkowsky, P. L. Callet and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2010.

[14] ITU-R, "Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures," in *International Telecommunication Union*, 2002.

[15] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak and U. Reiter, "Temporal development of quality of experience," in *Quality of Experience*, Springer International Publishing, 2014, pp. 133-147.

[16] D. E. Pearson, "Viewer response to time-varying video quality," in *Photonics West'98 Electronic Imaging International Society for Optics and Photonics*, 1998.

[17] A. Rehman and Z. Wang, "Perceptual experience of time-varying video quality," in *Quality of Multimedia Experience (QoMEX)*, 2013.