QUALITY ASSESSMENT OF A STEREO PAIR FORMED FROM DECODED AND SYNTHESIZED VIEWS USING OBJECTIVE METRICS

Philippe Hanhart and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG),
Ecole Polytechnique Fédérale de Lausanne (EPFL),
1015 Lausanne, Switzerland

ABSTRACT

When a stereo pair is formed from a decoded view and a synthesized view, it is unclear how the overall quality of the stereo pair should be assessed through objective quality metrics. In this paper, this problem is addressed considering a 3D video represented in the format of multiview video plus depth. The performance of different state-of-the-art 2D quality metrics is analyzed in terms of correlation with subjective perception of video quality. A set of subjective data collected through formal subjective evaluation tests is used as benchmark. Results show that the measured quality of the decoded view has the highest correlation with perceived quality. If the objective quality assessment is based on the measured quality of the synthesized view, it is suggested to use VIF, VQM, MS-SSIM, or SSIM since they significantly outperform other objective metrics, including PSNR.

***Index Terms* —** 3D, quality assessment, quality metric, asymmetric stereo, view synthesis.

1. INTRODUCTION

Despite the efforts of the scientific community in recent years, 3D video quality assessment is still an open challenge and there are no objective metrics which are widely recognized as reliable predictors of human 3D quality perception. The assessment of 3D quality is particularly challenging for mismatched or asymmetric stereoscopic videos, which have different strength and/or types of degradation between the left and right views. In general, the perceived quality of an asymmetric stereo pair is closer to the average quality of the two views. However, Stelmach *et al.* [1] have shown that, depending on the type of degradation and the difference of quality between the individual views, the 3D quality can be closer to the highest quality. Therefore, specific properties of the human visual system, such as binocular suppression (the masking of low-frequency content in one view by the sharp visual content in the other view), should be taken into account when building models that objectively quantify the 3D quality of a stereo pair.

In March 2011, a Call for Proposals (CfP) on 3D Video Coding Technology was issued by MPEG [2]. One of the objectives is to allow advanced processing of stereoscopic content to cope with varying display types and sizes, as well as different viewing preferences. For this application, a 2-view configuration is assumed, as illustrated in Figure 1. In this configuration, the decoded data, i.e., texture views and corresponding depth maps, is used to synthesize a virtual view at a selected position. The stereo pair displayed on the stereoscopic monitor consists of the decoded left or right view and the synthesized view. The displayed stereo pair is considered as asymmetric since one view contains only compression artifacts while the other view contains both compression and view synthesis artifacts. Due to the artifacts introduced by the view synthesis algorithm and the compression of the depth maps, it is expected that the individual quality of the virtual view is lower than that of the decoded view.

Bosc *et al.* [3] have shown that traditional objective metrics have a very low correlation with perceived quality when used for objective quality assessment of synthesized views. Therefore, for a stereo pair formed from a decoded view and a synthesized view, it is unclear wether objective metrics correlate well with perceived quality and which views should be taken into account: the decoded view, the synthesized view, or both views?

In our previous study [4], we had investigated the correlation between different PSNR-based metrics and the perceived quality of a stereo pair formed from a decoded view and a synthesized view. To evaluate the metrics performance, we used as ground truth subjective results collected during the evaluations of the MPEG CfP. Results showed that the PSNR of the decoded view had the highest correlation in terms of the Pearson correlation coefficient with perceived quality. Similar performance was achieved when using the average PSNR value of both views. On the other hand, the PSNR of the synthesized view had a significantly lower correlation with the subjective results.

In this study, we extend our analysis to other state-of-the-art 2D quality metrics, including perceptual based metrics, which might show different results compared to PSNR. The objective metrics are benchmarked following a similar methodology as in our previous study.

The paper is organized as follows. Section 2 provides an overview of the methodology followed in the evaluations to collect the subjective results used as benchmark in this study. The different objective metrics benchmarked in this study are defined in Section 3. In Section 4, the methodology used to evaluate the performance of the objective metrics is described. Results are shown and analyzed in Section 5. Conclusions and discussion on future work are presented in Section 6.

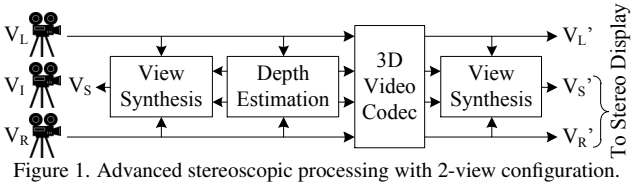

Figure 1. Advanced stereoscopic processing with 2-view configuration.

2. SUBJECTIVE QUALITY ASSESSMENT

The test material used in the MPEG CfP is composed of eight different contents encoded at four target bit rates. The contents are divided in two classes: Class A, with a spatial resolution of 1920×1088 pixels and a temporal resolution of 25 frames per seconds, and Class C, with 1024×768 pixels at 30 frames per second. All contents are 10 seconds long. All test sequences were stored as raw YUV video files. Twenty-two coding algorithms, submitted by the proponents, and two anchors were evaluated in the tests.

The evaluation was performed using a 46" Hyundai S465D polarized stereoscopic monitor with a native resolution of 1920×1080 pixels. Eighteen naive viewers evaluated the quality of each test sequence. The viewers were seated at a distance of about four times the height of the active part of the display. The laboratory setup had controlled lighting system to produce reliable and repeatable results. All subjects taking part in the evaluations underwent a screening to examine their visual acuity, color vision, and stereo vision.

The Double Stimulus Impairment Scale (DSIS) evaluation methodology was selected to perform the tests. Subjects were presented with pairs of video sequences (i.e., stimuli), where the first sequence was always a reference video (stimulus A) and the second, the video to be evaluated (stimulus B). Subjects were asked to rate the quality of each stimulus B, when compared to stimulus A. An 11-grade numerical categorical scale was used. The rating scale ranged from 0 (lowest quality) to 10 (highest quality). Before each test session, written instructions and a short explanation by a test operator were provided to the subjects. Also, a training session was run to show the graphical user interface, the rating sheets, and examples of processed video sequences. In each session, the stimulus pairs were presented in random orders, but never with the same video content in consecutive pairs.

The subjective results have been processed by first detecting and removing subjects whose scores appeared to deviate strongly from other scores in each test session. Then, the mean opinion score (MOS_i) was computed for each test sequence as the mean across the rates of the valid subjects, as well as associated standard deviation (σ_i) and 95% confidence interval, assuming a Student's t-distribution of the scores. Readers can refer to our previous paper [4] for more details.

3. OBJECTIVE QUALITY ASSESSMENT

In this study, the performance of the following objective metrics (OM) are assessed:

1. PSNR: Peak Signal-to-Noise Ratio,
2. SSIM: Structural Similarity Index [5],
3. MS-SSIM: Multi-Scale Structural Similarity Index [6],
4. VSNR: Visual Signal-to-Noise Ratio [7],
5. VIF: Visual Information Fidelity[1] [8],
6. WSNR: Weighted Signal-to-Noise Ratio[2] [9],
7. PSNR-HVS: PSNR Human Visual System [10],
8. PSNR-HVS-M: PSNR Human Visual System Masking [11],
9. VQM: Video Quality Metric[3] [12].

[1]Pixel domain version.

[2]This objective metric should not be confused with the weighted sum of the PSNR of the luma and chroma components.

[3]NTIA General Model, no calibration.

All above objective metrics, except for VQM, are computed on the luma component of each frame and the resulting values are averaged across the frames to produce a global index for the entire video sequence.

Most of the objective metrics, except for WSNR, VSNR, and VQM, were computed using our Video Quality Measurement Tool [13]. WSNR was computed using MeTriX MuX Visual Quality Assessment Package [14]. VSNR was obtained from its developer website [15]. VQM was obtained from the Institute for Telecommunication Sciences (ITS) website [16].

In the 2-view configuration, as considered in the MPEG CfP, a pair of cameras is used to produce the input views at the encoder side. At the decoder side, the displayed stereo pair is formed from the decoded right view and a synthesized view, located in-between the input views, as depicted in Figure 1. The baseline (inter-camera distance) of the displayed stereo pair is roughly equal to half of the baseline of the input stereo pair. For one Class A content and all Class C contents, the location of the synthesized view matched the location of a real view, called intermediate view, available in the original data (but not used by the encoder).

Five different objective video quality models are considered:

a) Quality of the decoded view, calculated between the decoded view and the original view: $\text{OM}(V_R^{'}, V_R)$
b) Quality of the intermediate view, calculated between the synthesized view at the decoder side and the intermediate view from the original data (when available): $\text{OM}(V_S^{'}, V_I)$
c) Quality of the synthesized view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side: $\text{OM}(V_S^{'}, V_S)$
d) Average quality of the decoded view and the intermediate view, computed as the mean value of a) and b)
e) Average quality of the decoded view and the synthesized view, computed as the mean value of a) and c)

4. PERFORMANCE INDEXES

The results of the subjective tests can be used as ground truth to evaluate how well the objective metrics estimate perceived quality. The result of execution of a particular objective metric and objective video quality model is a Video Quality Rating (VQR), which is expected to be the estimation of the MOS corresponding to a pair of video data. As compliant to the standard procedure for evaluating the performance of objective metrics [17], the following properties of the VQR estimation of MOSs are considered in this study: accuracy, monotonicity, and consistency.

First, a linear least squares regression is fitted to each [VQR, MOS] data set. The linear regression aligns the VQR range to the MOS range and allows removing any systematic offset, which may be present in the relationship between the objective and subjective data. This offset is irrelevant for the goal of metric performance comparison. At the same time, the linear regression avoids the risk of data over fitting, which may occur when considering non-linear regression. The linear regression is of the form:

$$\text{MOS}_p(\text{VQR}) = a \cdot \text{VQR} + b$$

Then, the Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) are computed between MOS_p and MOS to estimate the accuracy of the VQR. To estimate monotonicity and consistency, the Spearman rank order correlation coefficient (SCC) and the outlier ratio (OR), are computed between

MOS_p and MOS, respectively. Finally, these four estimators are averaged across the different contents.

The root-mean-square error (RMSE) and the outlier ratio (OR) are defined as follow:

$$\text{RMSE} = \sqrt{\frac{1}{(N-D)}\sum_{i=1}^{N}(\text{MOS}_i - \text{MOS}_{pi})^2}$$

$$\text{OR} = \frac{\text{total number of outliers}}{N}$$

$$\text{outlier: point for which } |\text{MOS}_i - \text{MOS}_{pi}| > 2\sigma_i$$

where N is the total number of points, D is the degree of freedom for the curve fitting (linear: $D = 2$), and σ_i is the standard deviation corresponding to MOS_i.

5. RESULTS

The accuracy, monotonicity, and consistency indexes of the objective video quality models, as defined in Section 4, are reported for each objective metric separately in Table 1. The objective metrics are ranked for each objective video quality model and the ranking number is specified below each performance index value.

It can be noticed that the PSNR of the intermediate view (PCC=0.5858, SCC=0.6234) has a significantly lower correlation with perceived quality than the PSNR of the synthesized view (PCC=0.6668, SCC=0.6797). PSNR-HVS and PSNR-HVS-M, which are based on PSNR, have a similar behavior. The difference between the intermediate and synthesized views is not significant for the other objective metrics.

For PSNR, PSNR-HVS, PSNR-HVS-M, WSNR, and VSNR, the objective video quality models that take into account the quality of the decoded view have a significantly higher correlation with perceived quality than the other objective video quality models. On the other hand, SSIM, MS-SSIM, VIF, and VQM have similar performance regardless the objective video quality model. A few hypotheses can be raised to explain these observations:

a) In terms of perceived quality, the higher quality of the decoded view, which does not contain view synthesis artifacts, tends to mask the lower quality of the synthesized view
b) Most of the considered objective metrics do not predict well perceived quality of synthesized views

Table 1. Accuracy, monotonicity, and consistency indexes of the objective metrics under consideration.

	Decoded	Intermediate	Synthesized	Decoded and intermediate	Decoded and synthesized	Decoded	Intermediate	Synthesized	Decoded and intermediate	Decoded and synthesized
	Pearson linear correlation coefficient (PCC)					Spearman rank order correlation coefficient (SCC)				
PSNR	**0.9200** 2	0.5858 9	0.6668 7	0.9028 4	0.8834 8	**0.9114** 7	0.6234 9	0.6797 6	0.8892 9	0.8762 8
SSIM	**0.9130** 7	0.8506 3	0.8460 4	0.8957 6	0.8987 3	**0.9080** 9	0.8655 4	0.8530 4	0.9077 5	0.9022 5
MS-SSIM	**0.9131** 6	0.8507 2	0.8495 3	0.8907 8	0.8977 5	**0.9177** 1	0.8675 3	0.8584 3	0.9094 3	0.9120 2
VSNR	0.9131 5	0.7188 5	0.7532 5	**0.9153** 2	0.8986 4	0.9118 6	0.7500 5	0.7642 5	**0.9201** 1	0.9040 4
VIF	**0.9094** 8	0.8510 1	0.8544 2	0.8930 7	0.9013 1	**0.9152** 3	0.8740 2	0.8744 2	0.9108 2	0.9128 1
WSNR	**0.9216** 1	0.6530 6	0.6729 6	0.9188 1	0.9002 2	**0.9166** 2	0.6687 6	0.6735 7	0.9082 4	0.8978 6
PSNR-HVS	**0.9194** 3	0.5913 8	0.6491 9	0.9012 5	0.8809 9	**0.9129** 5	0.6279 8	0.6585 9	0.8931 8	0.8747 9
PSNR-HVS-M	**0.9181** 4	0.5962 7	0.6507 8	0.9033 3	0.8836 7	**0.9139** 4	0.6324 7	0.6591 8	0.8964 7	0.8796 7
VQM	**0.8944** 9	0.8466 4	0.8599 1	0.8684 9	0.8874 6	**0.9105** 8	0.8765 1	0.8803 1	0.8966 6	0.9061 3
	Root-mean-square error (RMSE)					Outliers ratio (OR)				
PSNR	**0.9334** 2	1.7519 9	1.6795 7	1.0358 5	1.1108 7	**0.0099** 4	0.1013 9	0.1048 7	0.0278 4	0.0340 8
SSIM	**0.9713** 7	1.1942 2	1.2272 3	1.0440 7	1.0381 4	**0.0181** 6	0.0378 1	0.0405 3	0.0252 3	0.0299 4
MS-SSIM	**0.9877** 8	1.2328 4	1.2298 4	1.0981 8	1.0599 6	**0.0285** 8	0.0452 3	0.0511 4	0.0399 9	0.0359 9
VSNR	0.9664 6	1.3329 5	1.4171 5	**0.9450** 2	1.0199 2	**0.0120** 5	0.0587 5	0.0844 5	0.0146 1	0.0159 1
VIF	**0.9608** 5	1.1656 1	1.1823 2	1.0235 3	0.9918 1	**0.0184** 7	0.0378 2	0.0373 1	0.0289 7	0.0281 3
WSNR	**0.9306** 1	1.6325 6	1.6616 6	0.9449 1	1.0312 3	**0.0087** 3	0.0838 6	0.0993 6	0.0243 2	0.0271 2
PSNR-HVS	**0.9349** 3	1.7377 8	1.7413 9	1.0431 6	1.1265 9	**0.0060** 1	0.1002 7	0.1132 8	0.0278 5	0.0328 6
PSNR-HVS-M	**0.9429** 4	1.7258 7	1.7361 8	1.0307 4	1.1141 8	**0.0074** 2	0.1002 8	0.1132 9	0.0278 6	0.0328 7
VQM	**1.0237** 9	1.2092 3	1.1687 1	1.1212 9	1.0473 5	**0.0285** 9	0.0500 4	0.0395 2	0.0362 8	0.0308 5

The first hypothesis is in agreement with the results from Stelmach *et al.* [1]. The second hypothesis is in agreement with the results from Bosc *et al.* [3]. However, in their study, no compression artifacts were considered and the evaluation was performed with 2D still images only. It is also known that PSNR has good performance for compression artifacts but rather low performance for other types of degradation or when different types of degradations are combined. All these factors play an important role and should be further investigated to better understand how the viewer perceives quality of a stereo pair formed from a decoded view and a synthesized view and how it can be predicted using objective metrics. A similar study should be conducted using stereo pairs formed from two synthesized views to further investigate these hypotheses.

In general, the objective video quality model based on the quality of the decoded view has the highest correlation with perceived quality. In this case, all objective metrics have a high correlation (PCC$\geq$0.8944, SCC$\geq$0.9080) with perceived quality. If the objective quality assessment is based on the measured quality of the synthesized view, it is suggested to use VQM, VIF, MS-SSIM, or SSIM since these objective metrics have a significantly higher correlation with perceived quality (PCC$\geq$0.8460, SCC$\geq$0.8530). Taking into account both views increases correlation with perceived quality as opposed to using the synthesized (intermediate) view only.

6. CONCLUSION AND FUTURE WORK

In this paper, the correlation between different state-of-the-art objective 2D metrics and the perceived quality of a stereo pair formed from a decoded view and a synthesized view has been investigated. Results show that, in general, the measured quality of the decoded view has the highest correlation in terms of the Pearson correlation coefficient with perceived quality. Similar performance can be achieved when considering the average quality of both views. However, if the objective quality assessment is based on the measured quality of the synthesized view, it is suggested to use VIF, VQM, MS-SSIM, or SSIM since they significantly outperform other objective metrics, including PSNR. These four objective metrics have similar performance when using the decoded view, the synthesized view, and both views.

To better understand the masking effect between the decoded view and the synthesized view and further investigate the performance of the objective metrics in assessing quality of synthesized views, a similar study needs to be conducted with stereo pairs formed from two synthesized views.

7. ACKNOWLEDGEMENT

We would like to thank Panos Nasiopoulos and Mahsa T. Pourazad from University of British Columbia, and Kjell Brunnström, Kun Wang, and Börje Andrén from Acreo AB for providing subjective results for the 2-view configuration. This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET and Swiss SER project Quality of Experience in 3DTV.

8. REFERENCES

[1] L.B. Stelmach, W.J. Tam, D.V. Meegan, A. Vincent, and P. Corriveau, "Human Perception of Mismatched Stereoscopic 3D Inputs," in *International Conference on Image Processing*, September 2000, vol. 1, pp. 5–8.

[2] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on 3D Video Coding Technology," Doc. N12036, Geneva, Switzerland, November 2011.

[3] E. Bosc, M. Köppel, R. Pépion, M. Pressigout, L. Morin, P. Ndjiki-Nya, and P. Le Callet, "Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols?," in *International Conference on Image Processing*, 2011, pp. 2597–2600.

[4] P. Hanhart, F. De Simone, and T. Ebrahimi, "Quality Assessment of Asymmetric Stereo Pair Formed From Decoded and Synthesized Views," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, Yarra Valley, Australia, July 5-7, 2012.

[5] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[6] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, November 2003, vol. 2, pp. 1398–1402.

[7] D.M. Chandler and S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, September 2007.

[8] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, February 2006.

[9] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, April 2000.

[10] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, 2006.

[11] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2007.

[12] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," ITU-T Telecommunication Standardization Bureau, March 2004.

[13] http://mmspg.epfl.ch/vqmt/.

[14] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[15] http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html.

[16] http://vqm.its.bldrdoc.gov/.

[17] ITU-T Tutorial, "Objective perceptual assessment of video quality: Full reference television," ITU-T Telecommunication Standardization Bureau, 2004.