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

Philippe Hanhart and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG),
Ecole Polytechnique Fédérale de Lausanne (EPFL),
Station 11, 1015 Lausanne, Switzerland

ABSTRACT

When a stereo pair is formed from two synthesized views, it is unclear whether objective 2D quality metrics can provide a good estimation of the perceived quality. In this paper, this problem is addressed considering a 3D video represented in multiview video plus depth format. 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 some objective metrics, including PSNR, do not predict well perceived quality of synthesized views. On the other hand, metrics such as VIF, VQM, MS-SSIM, or SSIM have a high correlation with perceived quality.

1. INTRODUCTION

Understanding and measuring the effect of view synthesis on perceived quality, in conjunction with compression, is particularly important for multiview autostereoscopic displays, which usually synthesize N views from a limited number of input views, and stereoscopic displays that modify the baseline to adjust the depth perception based on viewing distance and viewing preferences. 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. Hewage *et al.* [1] have investigated objective quality assessment of 3D content represented in video plus depth (2D+Z) format using PSNR, SSIM, and VQM. The objective quality metrics were computed on the 2D video and on the rendered left and right 3D views. It was found that VQM had the highest correlation with perceived quality. The metrics showed lower correlation with perceived quality when using the average quality of the left and right 3D views than when using the quality of the 2D video. This effect was particularly strong for PSNR, where the correlation coefficient dropped from 0.81 to 0.74. However, Bosc *et al.* [2] have shown that traditional objective metrics, including PSNR, have a very low correlation with perceived quality when used for objective quality assessment of synthesized views. Nevertheless, in their study, no compression artifacts were considered and the evaluation was performed with 2D still images only. Therefore, for a stereo pair formed from two synthesized views, which were synthesized from a decoded 3D video represented in the multiview video plus depth (MVD) format, it is unclear whether objective 2D quality metrics can provide a good estimation of the perceived quality.

In March 2011, a Call for Proposals (CfP) on 3D Video Coding Technology was issued by MPEG [3]. One of the objectives was to support high-quality multiview autostereoscopic displays through generation of many high-quality views from a limited number of input views. For this application, a 3-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 set of virtual views at selected positions. The decoded and synthesized views are displayed on the multiview autostereoscopic monitor. The 3-view configuration was evaluated both on multiview autostereoscopic and stereoscopic displays. In the latter case, the displayed stereo pair is formed from two synthesized views, as specified in Table 1. More specifically, two different stereo pairs were

Table 1. Input views and displayed stereo pairs.

Seq. ID	Test Sequence	Test Class	Input views	Fixed stereo pair	Random stereo pair
S01	Poznan Hall2	A	7-6-5	6.125-5.875	-
S02	Poznan Street		5-4-3	4.125-3.875	-
S03	Undo Dancer		1-5-9	4.5-5.5	-
S04	GT Fly		9-5-1	5.5-4.5	-
S05	Kendo	C	1-3-5	2.75-3.25	2.25-2.75
S06	Balloons		1-3-5	2.75-3.25	4.375-4.875
S07	Lovebird1		4-6-8	5.75-6.25	4.0833-4.5833
S08	Newspaper		2-4-6	3.75-4.25	4.3333-4.8333

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.

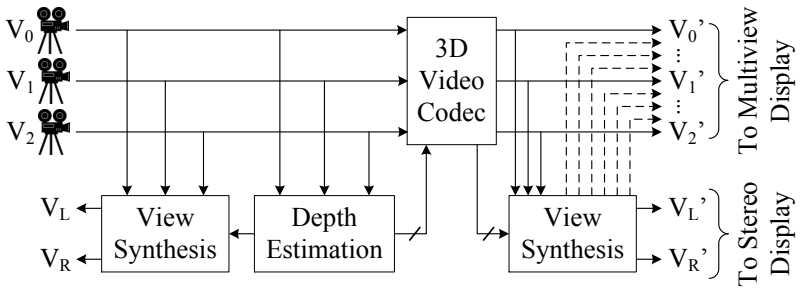

Fig. 1. Stereoscopic and autostereoscopic output with 3-view configuration.

evaluated: one referred to as fixed stereo pair, which is centered on the central decoded view, and one referred to as random stereo pair, which is located in-between two decoded views. The random pair was evaluated for Class C only.

In our previous study [4], we had investigated the correlation between different state-of-the-art 2D quality metrics, including perceptual 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 measured quality of the decoded view had the highest correlation in terms of the Pearson correlation coefficient with perceived quality. When the objective quality assessment was based on the measured quality of the synthesized view, results showed that VIF, VQM, MS-SSIM, and SSIM significantly outperformed other objective metrics. Two hypotheses were 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

In this paper, we report the results of a different problem, namely when a stereo pair is formed from two synthesized views, which might help us accepting or rejecting the hypotheses formulated in our previous paper. Following a similar methodology as in our previous studies, we benchmark the same objective metrics using a new set of stereoscopic videos and associated subjective quality scores.

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.

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. 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. Readers can refer to our previous paper [5] for more details.

In this paper, we used the mean opinion scores (MOS) that were computed by the MPEG test coordinator on a total of 36 naive viewers coming from three different laboratories [6]. Outlier detection was performed by the MPEG test coordinator according to the procedure adopted by the ITU Video Quality Experts Group (VQEG) for its Multimedia Project. Then, the MOS were computed for each test sequence as the mean across the rates of the valid subjects.

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. PSNR-HVS: PSNR Human Visual System [7],
3. PSNR-HVS-M: PSNR Human Visual System Masking [8],
4. WSNR: Weighted Signal-to-Noise Ratio[1] [9],
5. VSNR: Visual Signal-to-Noise Ratio [10],
6. SSIM: Structural Similarity Index [11],
7. MS-SSIM: Multi-Scale Structural Similarity Index [12],
8. VIF: Visual Information Fidelity[2] [13],
9. VQM: Video Quality Metric[3] [14].

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[4]. WSNR was computed using MeTriX MuX Visual Quality Assessment Package[5]. VSNR was obtained from its developer website[6]. VQM was obtained from the Institute for Telecommunication Sciences (ITS) website[7].

Three different objective video quality models are considered:

a) Quality of the left view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side: OM$(V_L^{'}, V_L)$
b) Quality of the right view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side: OM$(V_R^{'}, V_R)$
c) Average quality of both views, computed as the mean value of a) and b)

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 [15], the following properties of the VQR estimation of MOSs are considered in this study: accuracy and monotonicity.

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 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, the Spearman rank order correlation coefficient (SCC) is computed between MOS_p and MOS, respectively. Finally, these three estimators are averaged across the different contents.

The root-mean-square error is defined as follow:

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

where N is the total number of points and D is the degree of freedom for the curve fitting (linear: $D = 2$).

5. RESULTS

The accuracy and monotonicity indexes of the objective video quality models, as defined in Section 4, are reported for each objective metric separately in Table 2 and Table 3 for the fixed and random stereo pairs, respectively. The objective metrics are ranked for each objective video quality model and the ranking number is specified below each performance index value.

The fixed stereo pair is centered on the central decoded view and both views are equidistant from the central decoded view. Thus, both views should have the same amount of disocclusion and the same strength of view synthesis artifacts. The random stereo pair is located in-between two decoded views; one view of the stereo pair is always located closer to one of the decoded views than the other view of the stereo pair. Thus, we denote them as closer and farther views rather than left and right views. For example, for content S05 (see Table 1), view 2.75 is the closer view while view 2.25 is the farther view. The closer view has a lower amount of disocclusion than the farther view. Thus, the closer view should contain less view synthesis artifacts than the farther view. However, for the random stereo pair, there is no significant difference between the results for the closer and farther views ($\max|\Delta\text{PCC}| = 0.0146$, $\max|\Delta\text{SCC}| = 0.0114$). In general, the objective video quality model based on the average quality of

[1] This objective metric should not be confused with the weighted sum of the PSNR of the luma and chroma components.
[2] Pixel domain version.
[3] NTIA General Model, no calibration.
[4] http://mmspg.epfl.ch/vqmt/
[5] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
[6] http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html
[7] http://vqm.its.bldrdoc.gov/

Table 2. Fixed stereo pair: accuracy and monotonicity indexes of the objective metrics under consideration.

	Pearson linear correlation coefficient (PCC)			Spearman rank order correlation coefficient (SCC)			Root-mean-square error (RMSE)		
	Left view	Right view	Average	Left view	Right view	Average	Left view	Right view	Average
PSNR	0.78919	0.80849	**0.80869**	0.79579	0.80959	**0.80969**	1.35819	1.30539	**1.30159**
PSNR-HVS	0.79958	0.81908	**0.81908**	0.80388	0.81678	**0.81798**	1.33048	1.27468	**1.27258**
PSNR-HVS-M	0.80167	0.82087	**0.82107**	0.80437	0.81757	**0.81877**	1.32747	1.27117	**1.26897**
WSNR	0.83736	**0.85876**	0.85866	0.83866	0.85266	**0.85366**	1.20876	**1.13106**	1.13186
VSNR	0.90505	**0.92811**	0.92671	0.91685	**0.93393**	0.93245	0.93134	**0.82741**	0.83992
SSIM	0.91893	0.92054	**0.92153**	0.92954	0.93115	**0.93244**	0.88573	0.87694	**0.87214**
MS-SSIM	**0.90744**	0.90465	0.90735	0.93741	0.93592	**0.93881**	**0.94295**	0.95745	0.94495
VIF	0.92141	0.92412	**0.92452**	0.93662	0.93621	**0.93822**	0.85631	**0.83762**	0.83771
VQM	0.91962	**0.92103**	0.92084	0.93353	0.93184	**0.93373**	0.86132	**0.85283**	0.85423

Table 3. Random stereo pair: accuracy and monotonicity indexes of the objective metrics under consideration.

	Pearson linear correlation coefficient (PCC)			Spearman rank order correlation coefficient (SCC)			Root-mean-square error (RMSE)		
	Closer view	Farther view	Average	Closer view	Farther view	Average	Closer view	Farther view	Average
PSNR	0.70779	0.70829	**0.71229**	0.73909	0.74009	**0.74159**	1.59039	1.60419	**1.58809**
PSNR-HVS	0.72168	0.72168	**0.72658**	0.74428	0.74527	**0.74808**	1.55998	1.57548	**1.55648**
PSNR-HVS-M	0.72567	0.72627	**0.73097**	0.74567	0.74528	**0.74977**	1.55427	1.56637	**1.54847**
WSNR	0.75696	0.75876	**0.76336**	0.77356	0.76526	**0.77846**	1.47216	1.47776	**1.46096**
VSNR	0.83685	0.85145	**0.85175**	0.84955	0.84195	**0.85695**	1.16375	1.16745	**1.14365**
SSIM	0.93073	**0.94042**	0.93842	0.93384	**0.94522**	0.94273	0.84523	**0.79492**	0.80562
MS-SSIM	0.90924	0.90504	**0.90994**	0.93383	0.93264	**0.93694**	0.97114	0.99454	**0.97024**
VIF	0.93731	0.94251	**0.94341**	0.94422	0.95001	**0.95111**	0.80981	0.77271	**0.76931**
VQM	0.93142	0.92943	**0.93243**	**0.94661**	0.93923	0.94532	0.83642	0.84483	**0.82793**

both views has the highest correlation with perceived quality, but the difference between the models is not significant ($\max|\Delta\text{PCC}| = 0.0231$, $\max|\Delta\text{SCC}| = 0.0171$).

For stereo pairs formed from a decoded view and a synthesized view [4], the SNR-based metrics (PSNR, PSNR-HVS, PSNR-HVS-M, WSNR, and VSNR) had significantly lower correlation with perceived quality than the perceptual metrics (VIF, VQM, SSIM, and MS-SSIM) when using the synthesized view. In this study, a similar behavior is observed on the three objective video quality models for stereo pairs formed from two synthesized views. The results reported in this paper show that PSNR, PSNR-HVS, PSNR-HVS-M, and WSNR have a significantly lower correlation with perceived quality than VIF, VQM, SSIM, and MS-SSIM. The difference is particularly strong for the random stereo pair between SNR-based metrics (PCC ≤ 0.7633 and SCC ≤ 0.7784) and perceptual metrics (PCC ≥ 0.9050 and SCC ≥ 0.9326). In this case, PSNR (PCC ≤ 0.7122, SCC ≤ 0.7415) has a significantly lower correlation with perceived quality compared to VIF (PCC ≥ 0.9373, SCC ≥ 0.9442). For the fixed stereo pair, all perceptual metrics (PCC ≥ 0.9046 and SCC ≥ 0.9295) outperform PSNR (PCC ≤ 0.8086 and SCC ≤ 0.8096).

For stereo pairs formed from a decoded view and a synthesized view [4], the maximum absolute difference, calculated between the different objective video quality models, of PCC and SCC values are reported for each objective metric separately in Table 4. Only the quality of the decoded view, the quality of the synthesized view, and the average quality of the decoded view and the synthesized view are considered. The difference between the objective video quality models is about four times higher for PSNR, PSNR-HVS, PSNR-HVS-M, and WSNR ($\max|\Delta\text{PCC}| \geq 0.2487$, $\max|\Delta\text{SCC}| \geq 0.2317$) than for the perceptual metrics ($\max|\Delta\text{PCC}| \leq 0.0670$, $\max|\Delta\text{SCC}| \leq 0.0593$). There is a significant difference in performance between the different objective video quality models for these SNR-based metrics. However, the perceptual metrics have similar performance regardless the objective video quality model.

Table 4. Difference between objective video quality models.

	$\max\|\Delta\text{PCC}\|$	$\max\|\Delta\text{SCC}\|$
PSNR	0.2532	0.2317
PSNR-HVS	0.2703	0.2544
PSNR-HVS-M	0.2674	0.2548
WSNR	0.2487	0.2431
VSRN	0.1599	0.1476
SSIM	0.0670	0.0550
MS-SSIM	0.0636	0.0593
VIF	0.0550	0.0408
VQM	0.0345	0.0302

The results obtained for stereo pairs formed from two synthesized views lead to similar conclusion than the results obtained for stereo pairs formed from a decoded view and a synthesized view. These results indicate that some objective metrics do not predict well perceived quality of synthesized views and we must accept our second hypothesis. This conclusion is in line with the results from Bosc *et al.* [2].

Let's now consider only the objective metrics that have a high correlation with perceived quality of synthesized views, namely VIF, VQM, SSIM, and MS-SSIM. If there was a masking effect between the decoded view and the synthesized view in our previous study [4], we should have observed a significant difference between the objective video quality model based on the quality of the decoded view and the objective video quality model based on the quality of the synthesized view. However, these metrics have similar performance regardless the objective video quality model. These results indicate that there is no significant masking effect between a decoded view and a synthesized view and we must reject our first hypothesis.

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 two synthesized views has been investigated. Results show that PSNR, PSNR-HVS, PSNR-HVS-M, and WSNR have a significantly lower correlation with perceived quality than VIF, VQM, SSIM, and MS-SSIM. For a stereo pair formed from a decoded view and a synthesized view, previous results showed a similar behavior when the objective quality assessment was based on the measured quality of the synthesized view. On the other hand, no significant difference was observed between the metrics when the objective quality assessment was based on the measured quality of the decoded view. From these observations, we conclude that some objective metrics do not predict well perceived quality of synthesized views and that there is no significant masking effect between a decoded view and a synthesized view.

To extend our work, 3D metrics should also be evaluated for the same target application in future investigations. To better understand the limitations of the objective quality metrics, an analysis of the resolving power of the metrics will be conducted.

7. REFERENCES

[1] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, S. Villette, and A.M. Kondoz, "Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304–318, April 2009.

[2] 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.

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

[4] P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Decoded and Synthesized Views Using Objective Metrics," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Zurich, Switzerland, October 15-17, 2012.

[5] 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.

[6] ISO/IEC JTC1/SC29/WG11, "Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology," Doc. N12347, Geneva, Switzerland, November 2011.

[7] 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*, January 2006.

[8] 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.

[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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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