CROWD-BASED QUALITY ASSESSMENT OF MULTIVIEW VIDEO PLUS DEPTH CODING

Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

ABSTRACT

Crowdsourcing is becoming a popular cost effective alternative to lab-based evaluations for subjective quality assessment. However, crowd-based evaluations are constrained by the limited availability of display devices used by typical online workers, which makes the evaluation of 3D content a challenging task. In this paper, we investigate two possible approaches to crowd-based quality assessment of multiview video plus depth (MVD) content on 2D displays: by using a virtual view and by using a free-viewpoint video, which corresponds to a smooth camera motion during a time freeze. We conducted the crowdsourcing experiments using seven MVD sequences encoded at different bit rates with the upcoming 3D-AVC video coding standard. The results demonstrate high correlation with subjective evaluations performed using a stereoscopic monitor in a controlled laboratory environment. The analysis shows no statistically significant difference between the two approaches.

Index Terms— 3D, multiview video plus depth, compression, subjective quality assessment, crowdsourcing, free-viewpoint video.

1. INTRODUCTION

Continuous improvements in three-dimensional (3D) video technologies demand for effective assessment approaches of 3D video compression algorithms. However, despite numerous recent efforts, 3D quality assessment is still an open challenge, since there is no metric that is widely recognized as a reliable predictor of perceived 3D quality. Therefore, subjective evaluation remains the most reliable assessment method, even though it is time consuming and expensive.

To reduce the costs of subjective evaluations and also to consider more practical environments, researchers are investigating crowdsourcing platforms, which allow employing workers online from around the world. The authors of [1] provide a comprehensive overview of crowdsourcing approaches for subjective evaluations of image and video content and [2] discusses and compares the corresponding existing implementation frameworks. Both works also discuss issues and limitations of crowdsourcing in the context of subjective evaluations. One of the constraints is the limited variety of display devices used by online workers. Due to this limitation, for example, a direct evaluation of 3D content is impossible. Therefore, it is necessary to use alternative 2D representations of 3D content in crowdsourcing evaluations.

One simple approach to representing 3D content on a 2D display is to play only one view of the 3D content. The intended depth perception would be lost with this approach, but it may be enough for the evaluation of compression, as many compression artifacts would still be visible even in one view. An alternative approach was proposed by Bosc *et al.* [3], which is a new protocol for evaluation of depth map compression algorithms on 2D displays. The authors used a free-viewpoint video (FVV) [4] sequence corresponding to a smooth camera motion during a time freeze, which was generated from fifty intermediate views in-between the left and right views of the original content. The resulting effect is similar to the 'bullet time' visual effect used in such movies like "The Matrix". The intermediate views were generated through depth-image-based rendering (DIBR) [5] using the decoded depth maps and original texture views. The FVV sequence can then be displayed on a regular 2D monitor and, more importantly, it still retains a depth perception without the aid of any special glasses, thanks to the motion parallax, which is known to be a strong monocular depth cue [6].

In this paper, we use virtual view synthesized from the MVD content (referred to as 'mono' in the paper) and FVV sequence corresponding to a smooth camera motion during a time freeze (referred to as 'sweep') to investigate both protocols for the quality assessment of multiview plus depth (MVD) content on 2D displays. Seven MVD sequences were encoded at different bit rates using the upcoming 3D-AVC video coding standard. A reference ground truth was obtained via a subjective evaluation of stereo pairs on a stereoscopic monitor in a laboratory environment. Then, two 'mono' and 'sweep' 2D representations were generated for each bit rate and evaluated in a crowdsourcing environment. To evaluate the suitability of crowd-based quality assessment of MVD coding, the results of the crowd-based evaluations were compared to the ground truth results of the lab-based evaluations.

The paper is organized as follows. Section 2 presents the details of the subjective evaluations. Section 3 describes the statistical tools used to analyze the results. Results are presented and analyzed in Sec. 4. Section 5 concludes the paper.

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), EC funded Network of Excellence VideoSense, and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET.

Table 1: Multiview video plus depth contents used in the experiments.

Sequence	Characteristics			Encoding			Lab			Mono			Sweep		
	Resolution	Frames	FPS	QP settings R4, R3, R2, R1	Depth resolution	Input views	Views	Frames	FPS	View	Frames	FPS	Views	Frame	FPS
Poznan Hall2	1920×1088	0-199	25	26, 31, 36, 41	Quarter	7-6-5	6.25-5.75	0-199	25	5.75	0-199	25	7:-0.02:5	199	15
Poznan Street	1920×1088	0-249	25	26, 31, 36, 41	Quarter	5-4-3	4.25-3.75	0-249	25	3.75	0-249	25	5:-0.02:3	249	15
Undo Dancer	1920×1088	0-249	25	31, 38, 41, 46	Full	1-5-9	4-6	0-249	25	6	0-249	25	1:0.08:9	249	15
GT Fly	1920×1088	0-249	25	26, 31, 36, 41	Quarter	9-5-1	6-4	0-249	25	4	0-249	25	9:-0.08:1	156	15
Kendo	1024×768	0-299	30	26, 31, 36, 41	Quarter	1-3-5	2.5-3.5	0-299	30	3.5	0-299	30	1:0.04:5	0	15
Balloons	1024×768	0-299	30	26, 31, 36, 41	Quarter	1-3-5	2.5-3.5	0-299	30	3.5	0-299	30	1:0.04:5	0	15
Newspaper	1024×768	0-299	30	26, 31, 36, 41	Quarter	2-4-6	3.5-4.5	0-299	30	4.5	0-299	30	2:0.04:6	0	15

2. METHODOLOGY

The single-stimulus (SS) methodology [7] was chosen as this methodology was selected in [3]. A five-grade quality scale (1: *Bad*; 2: *Poor*; 3: *Fair*; 4: *Good*; 5: *Excellent*) was used. The subjects were asked to judge the overall quality of the evaluated video sequence.

Four dummy video sequences (one with high quality, one with low quality, and two of mid quality), whose scores were not included in the results, were included at the beginning of the test session to stabilize the subjects' ratings. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each subject, whereas the same content was never shown consecutively.

2.1. Dataset

Seven MVD sequences were used in the experiments, with different visual characteristics, resolutions, and frame rates (see Table 1). All sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bits per sample. The sequences were compressed with 3D-AVC using 3D-ATM v9.0 [8] under the conditions defined in [9]. For each sequence, 5 stimuli were generated, 1 from the original data, and 4 from the decoded data, resulting in a total of 28 test stimuli. Five training samples were generated from the *Poznan CarPark* sequence, which was not used in the tests. Their quality was manually selected by expert viewers so that they represent all grades of the rating scale.

2.2. Lab-based evaluation

The stereo pairs were synthesized from the decoded data using VSRS-1D-Fast v8.0 [10], according to the parameters given in Table 1. The stereo pairs were displayed on a full HD 46" Hyundai S465D polarized stereoscopic monitor. The monitor was calibrated using an X-Rite i1Display Pro color calibration device according to the following profile: sRGB gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level. The test room was equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of maximum screen luminance.

The experiment involved up to three subjects assessing the test materials. Subjects were seated in a row perpendicular to the center of the monitor, at a distance of 3.2 times the picture height, as suggested in [7]. A total of 22 naïve subjects took part in the experiment. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively.

Before the experiment, oral instructions were provided to the subjects to explain their tasks. Additionally, a training session using the five training samples was organized to allow subjects to familiarize with the assessment procedure.

2.3. Crowd-based evaluation

Since no video player is capable of decoding 3D-AVC bit streams and synthesizing virtual views in real time, the video sequences were generated offline. The video sequences were synthesized from the decoded data using VSRS-1D-Fast v8.0 [10], according to the parameters given in Table 1. For the 'mono' representation, the right view of the stereo pair was used. For the 'sweep' representation, the FVV sequences were generated from a stack of 100 frames (at 15 fps), which was built from 50 intermediate views in-between the left and right views of the original content, according to [3]. One key frame, which maximizes the amount of depth, was selected as the freeze point for each content.

The sequences were encoded with H.264/MPEG-4 AVC High Profile, since transmitting uncompressed video data to remote workers is impractical, especially for full HD content. Original full HD sequences of 25 fps were compressed at 20 Mbit/s, which is commonly considered as perceptually transparent quality for video broadcasting. For other sequences, the bit rate was set proportionally. A two-pass encoding was used and the deblocking filter was disabled to preserve the original blockiness due to 3D-AVC at low bit rates. Expert viewers evaluated the quality of the compressed sequences as visually lossless. The full HD sequences were cropped to 1856×1016 pixels such that workers were able to see the whole video in the web browser on a WUXGA (1920×1200) monitor. To display the video sequences and collect individual scores, the QualityCrowd 2 framework [11] was used.

The experiments were conducted at EPFL in an uncontrolled computer lab, as it is relatively difficult to find workers equipped with a full HD monitor and because of the relatively large amount of transmitted video data (up to 670 MB). Therefore, the workers demographic was limited to EPFL stu-

dents. Also, no financial compensation was provided to the workers. Each worker evaluated all test stimuli. The same set of workers took part in the 'mono' (20 subjects) and 'sweep' (21 subjects) experiments. However, half of the subjects took part in the 'mono' experiment first, while the other half started with the 'sweep' experiment. To minimized memory effects, subjects took a break between the two experiments.

Before the experiments, short written instructions were provided to the workers to explain their tasks. Additionally, three training samples, representative of *Excellent*, *Fair*, and *Bad* quality, were displayed to familiarize workers with the assessment procedure.

3. STATISTICAL ANALYSIS

The subjective scores were processed by first detecting and removing subjects whose scores deviated strongly from others (for each experiment independently), as per the outlier detection procedure defined in [12], resulting in zero detected outliers. Then, the mean opinion score (MOS) was computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% confidence interval (CI), assuming a Student's *t*-distribution of the scores.

3.1. Performance indexes

To be compliant with the standard procedure for comparing MOS values of different experiments [13], a regression was fitted to each [MOS^{crowd},MOS^{lab}] data set, using linear and cubic fitting, with the constraint that the function is monotonic on the interval of observed values. The Pearson linear correlation coefficient (PLCC) and root-mean-square error (RMSE) were computed to estimate accuracy. The Spearman rank order correlation coefficient (SROCC) and outlier ratio (OR) were computed to estimate monotonicity and consistency, respectively. A sample i was considered as outlier if $\left|MOS_i^{lab} - MOS_i^p\right| > CI_i^{lab} + CI_i^{crowd}$, where CI_i^{lab} and CI_i^{crowd} are the 95% CIs corresponding to MOS_i^{lab} and MOS_i^{crowd}, respectively, and MOS_i^p is the fitted MOS.

3.2. Estimation errors

To determine whether the difference between two sets of scores corresponding to the same decoded 3D data evaluated in two different experiments is statistically significant, a multiple comparison test based on ANOVA was performed at a 5% significance level on the raw scores. The percentage of *Correct estimation*, *Underestimation*, and *Overestimation* were recorded from all data points.

3.3. Classification errors

In [14], it is recommended to determine the classification errors of an objective metric to evaluate its effectiveness. A classification error is made when the subjective test and the objective metric lead to different conclusions on a pair of data points. In [15], this definition was extended to compare the results of two subjective tests. A classification error is made when the two subjective tests lead to different conclusions on a pair of data points. Three types of errors can occur: *False tie* (the least offensive error), *False differentiation*, and *False ranking* (the most offensive error).

To determine whether the difference between two sets of scores corresponding to a pair of decoded 3D data evaluated in the same experiment is statistically significant, a multiple comparison test was performed similarly to Sec. 3.2.

4. RESULTS AND DISCUSSIONS

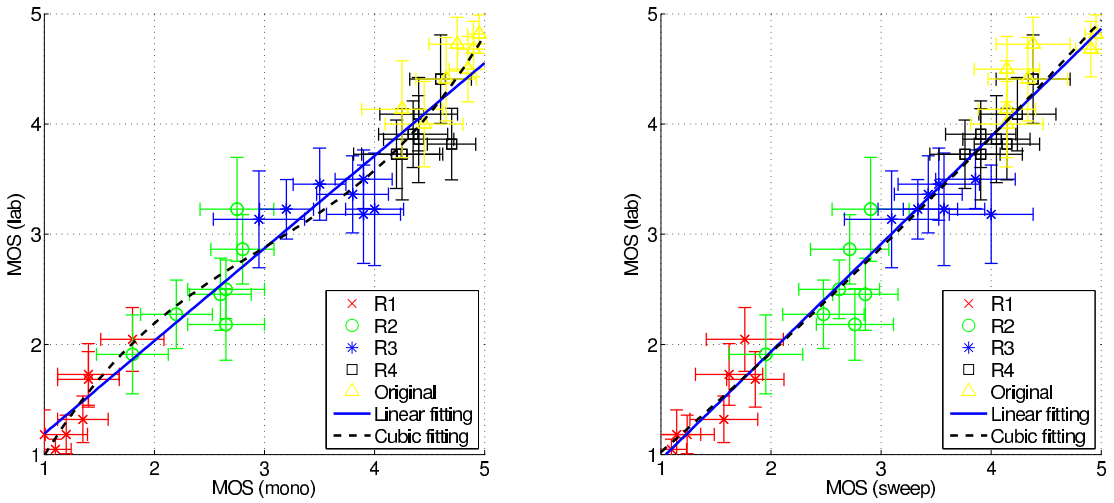
Figure 1(a) and Figure 1(b) show the results obtained for 'mono' and 'sweep' experiments respectively, with x-axis corresponding to the crowdsourcing data and y-axis to the data from the lab experiment. The horizontal and vertical error bars are the confidence intervals of the respective experiments. To illustrate the trends of the data points, linear and cubic regressions fitted to each data set are also shown.

Ideally, all points would be on a 45° line if the MOS values for each condition were the same between two experiments. While, in the figures, some points lie above the line and others lie below, no significant systematic offset can be observed among MOS values of the compared experiments. It means that, overall, crowdsourcing workers scored closely to the lab experiment.

In case of 'mono' vs. lab, the slope of the linear regression is a little smaller than 45° (see Figure 1(a)), indicating that workers scored more pessimistically on lower quality stimuli, which is probably due to using H.264/MPEG-4 AVC for encoding of the test stimuli instead of showing the original uncompressed data. More optimistic scores for high quality stimuli might be due to the uncalibrated monitors and uncontrolled lighting conditions. In Figure 1(b), the linear regression is very similar to a 45° line. The cubic regression for 'sweep' vs. lab is very close to a straight line, which means the relationship between 'sweep' and lab is more linear than between 'mono' and lab.

Table 2 reports the performance indexes. Results show that there is a very strong correlation between crowd-based and lab-based evaluations, as the correlation indexes are above 0.97, which is similar to the correlation between different laboratories conducting the same experiment on stereoscopic monitors [16]. The PLCC, RMSE, and OR indexes are slightly better for 'sweep' than 'mono' when no fitting or linear fitting are considered. However, there are no statistically significant differences between the performance indexes computed for 'mono' and 'sweep'.

Regardless of the fitting applied to the data sets, both crowd-based evaluations were able to correctly estimate the results of the lab-based evaluation with a *Correct estimation* of 100%, whereas the *Underestimation* and *Overestimation* were always null.



(a) Lab versus mono. (b) Lab versus sweep.

Fig. 1: Comparison of MOS values obtained in the different experiments.

Table 2: Performance indexes.

Fitting	Mono				Sweep			
	PLCC	SROCC	RMSE	OR	PLCC	SROCC	RMSE	OR
None	0.9750	0.9753	0.3697	2.86%	0.9761	0.9717	0.2629	0%
Linear	0.9750	0.9753	0.2495	2.86%	0.9761	0.9717	0.2440	0%
Cubic	0.9798	0.9753	0.2243	0%	0.9764	0.9717	0.2422	0%

Table 3: Classification errors.

Fitting	Mono				Sweep			
	Correct decision	*False ranking*	*False differentiation*	*False tie*	*Correct decision*	*False ranking*	*False differentiation*	*False tie*
None	86.05%	0.00%	6.89%	7.06%	87.90%	0.00%	4.54%	7.56%
Linear	84.54%	0.00%	5.71%	9.75%	87.90%	0.00%	4.54%	7.56%
Cubic	87.06%	0.00%	5.88%	7.06%	88.24%	0.00%	4.37%	7.39%

Table 3 reports the classifications errors. More than 84% of all possible distinct pairs of decoded 3D data lead to the same conclusion in crowd-based evaluations when compared to the lab-based evaluation. Moreover, *False Ranking* never occurs. Results for *False differentiation* show a slight advantage for 'sweep', but differences are not significant.

Displaying one view of the 3D content allows judging spatial and temporal impairments, whereas depth impairments are difficult to evaluate. Nevertheless, some depth impairments may be visible when considering a virtual view that is synthesized from video and depth data. The FFV sequence is better to judge depth impairments, but temporal impairments cannot be evaluated. However, the selection of the key frame may impact the perceived quality, as the strength of the impairments typically varies in time.

In our experiments, 2D impairments were mostly visible in the test material, even though depth maps were also compressed, and the strength of the spatial impairments was similar across time. Therefore, it is reasonable to have high correlation with ground truth results in 'mono' and 'sweep'. However, if the test material mostly contains depth impairments, the 'sweep' methodology is expected to be more suitable.

5. CONCLUSION

In this paper, we investigated two possible approaches to crowd-based quality assessment of multiview video plus depth content on 2D displays: by using a virtual view and by using a free-viewpoint video corresponding to a smooth camera motion during a time freeze. We conducted the corresponding crowdsourcing experiments using seven MVD sequences encoded at different bit rates with the upcoming 3D-AVC video coding standard. The crowdsourcing results showed high correlation with ground truth results obtained in a subjective evaluation performed on a stereoscopic monitor in a laboratory environment. No statistically significant differences between the two approaches were found.

6. REFERENCES

[1] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[2] T. Hossfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment," in *International Workshop on Multimedia Signal Processing (MMSP)*, Jakarta, Indonesia, September 2014.

[3] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for Free-viewpoint video sequences synthesized from decompressed depth data," in *5th International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Wörthersee, Austria, July 2013.

[4] A. Smolic, K. Mueller, P. Merkle, and A. Vetro, "Development of a new MPEG standard for advanced 3D video applications," in *6th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Salzburg, Germany, September 2009.

[5] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Stereoscopic Displays and Virtual Reality Systems*, May 2004, vol. 5291 of *Proc. SPIE*.

[6] B. Rogers and M. Graham, "Motion parallax as an independent cue for depth perception," *Perception*, vol. 8, no. 2, pp. 125–34, 1979.

[7] ITU-R BT.2021, "Subjective methods for the assessment of stereoscopic 3DTV systems," International Telecommunication Union, August 2012.

[8] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, "3D-AVC Test Model 9," Doc. JCT3V-G1003, San Jose, USA, January 2014.

[9] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, "3D Video Subjective Quality Assessment Test Plan," Doc. JCT3V-F1011, Geneva, Switzerland, October 2013.

[10] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, "Test Model 7 of 3D-HEVC and MV-HEVC," Doc. JCT3V-G1005, San Jose, USA, January 2014.

[11] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd - A framework for crowd-based quality evaluation," in *Picture Coding Symposium (PCS)*, Kraków, Poland, May 2012.

[12] F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *Journal of Visual Communication and Image Representation*, vol. 22, no. 8, pp. 734–748, 2011.

[13] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.

[14] ITU-T J.149, "Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)," International Telecommunication Union, March 2004.

[15] P. Hanhart and T. Ebrahimi, "On the evaluation of 3D codecs on multiview autostereoscopic display," in *4th IEEE International Workshop on Hot Topics in 3D (Hot3D)*, San Jose, California, USA, July 2013.

[16] M. Barkowsky, J. Li, T. Han, S. Youn, J. Ok, C. Lee, C. Hedberg, I. V. Ananth, K. Wang, K. Brunnström, and P. Le Callet, "Towards standardized 3DTV QoE assessment: cross-lab study on display technology and viewing environment parameters," in *Stereoscopic Displays and Applications XXIV*, March 2013, vol. 8648 of *Proc. SPIE*.