CROSS-LAB SUBJECTIVE EVALUATION OF THE MVC+D AND 3D-AVC 3D VIDEO CODING STANDARDS

Philippe Hanhart[1]*, Naeem Ramzan*[2]*, Vittorio Baroncini*[3]*, and Touradj Ebrahimi*[1]

[1] Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland
[2] School of Computing, University of West of Scotland, Paisley, Scotland
[3] Fondazione Ugo Bordoni, Rome, Italy

ABSTRACT

Consistent and imitable subjective measurement 3D video quality assessment is investigated for evaluating 3D service parameters and as an essential criterion towards the development of objective models. This paper analyzes the results obtained from three test laboratories to evaluate the performance of the MVC+D and 3D-AVC 3D video coding standards using identical video content and following similar methodologies and instructions. The correlation between laboratories was investigated using similar analysis to benchmarking of objective metrics. Analyses show that test laboratories employing different displays and different subjects could still produce highly correlated results, as they follow similar guidelines to carry out assessments.

Index Terms— Subjective evaluation, viewing environment, cross-lab validation, MVC+D, 3D-AVC

1. INTRODUCTION

Consistent and imitable subjective measurement 3D video quality assessment is investigated for evaluating 3D service parameters and as an essential criterion towards the development of objective models. Quality assessment of 3D video is identified to range over numerous psychophysical extents, e.g., picture excellence, depth perception, and visual comfort, which may lead to higher level insights, e.g., visual experience and naturalness.

An important factor in subjective quality assessment experiments is the viewing conditions and it can be greatly influenced in the case of 3D video where the perception of depth is an additional factor when compared to 2D video. Furthermore, selecting and calibrating the display is very crucial in 3D video as it has a significant effect on the perceived brightness and overall quality, especially when more than one lab is involved in subjective evaluation. It is an interesting and challenging task to conduct the 3D video quality assessment in different labs and attempt to simulate the same conditions. It helps not only to measure the video quality accurately and precisely, but also gives us foundation to define the objective metrics for 3D video.

Perkis *et al.* [1] performed cross-lab video quality assessment of 3D video to address various issues regarding certification of multimedia quality assessment. They evaluated two test scenarios, namely, a 2-view input configuration, on stereoscopic display, and a 3-view input configuration, on both auto-stereoscopic as well as stereoscopic display. However, in any single scenario, only two laboratories results were considered for cross validation.

Recently, Barkowsky *et al.* [2] have studied cross-lab 3DTV quality assessment method with a main focus on defining the effect of different lab conditions like passive polarized displays, active shutter displays, viewing distance, number of parallel viewers, and voting device.

In Nov. 2013, the Joint Collaborative Team on 3D Video Coding (JCT-3V) issued a test plan for 3D video subjective assessment [3] to evaluate the performance of two amendments of the H.264/AVC video coding standard, namely

1. Multiview video coding (MVC) compatible extension including depth (MVC+D) [4]: the main target of this extension is to enable 3D enhancements while maintaining MVC stereo compatibility.
2. AVC compatible video-plus-depth extension (3D-AVC) [5]: the main aims for higher compression efficiency by jointly compressing texture and depth data.

To analyze and compare the performance of the proposed technologies, a formal subjective quality evaluation was carried out, and a set of test video sequences, encoded with the proposed technologies, was produced. Three laboratories took part in the evaluation campaign of this test material: at EPFL in Switzerland, UWS in Scotland, and FUB in Italy.

This paper analyzes the results obtained from three subjective experiments on the aforementioned coding techniques using identical video content and following similar methodologies and instructions. Cross-laboratory analysis is performed to find out whether or not consistent results can be obtained. The correlation between laboratories was inves-

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1) 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	Resolution	Frame rate	Frames	QP settings	Depth resolution	Input views	SS stereo pair	OS stereo pair
Poznan Hall2	1920×1088	25 fps	$0-199$	$26, 31, 36, 41$	Quarter	$7-6-5$	$6.25-5.75$	$6-5.5$
Poznan Street	1920×1088	25 fps	$0-249$	$26, 31, 36, 41$	Quarter	$5-4-3$	$4.25-3.75$	$4-3.5$
Undo Dancer	1920×1088	25 fps	$0-249$	$31, 38, 41, 46$	Full	$1-5-9$	$4-6$	$5-7$
GT Fly	1920×1088	25 fps	$0-249$	$26, 31, 36, 41$	Quarter	$9-5-1$	$6-4$	$7-5$

tigated using similar analysis to benchmarking of objective metrics. In particular, the Pearson and Spearman correlation coefficients, root-mean-square error, and outlier ratio were computed between the sets of subjective scores. Additionally, the estimation and classification errors were considered to estimate the differences when comparing subjective scores between two laboratories. These analyses show that laboratories employing different displays and different subjects could still produce highly correlated results, as they follow similar guidelines to carry out the evaluations. This confirms that the participating laboratories have high correlation to conduct subjective evaluation.

The remainder of the paper is organized as follows. Section 2 describes the test design and methodologies along with different lab environments. Section 3 explains the statistical analysis performed on the results obtained from different labs. Detailed test results obtained in the cross validation of data and relevant analyses are presented in Section 4. Finally, concluding remarks are provided in Section 5.

2. METHODOLOGY

2.1. Dataset

Four multiview video plus depth (MVD) sequences, with different visual characteristics, were used in the experiments (see Table 1). The encoded views used in the experiments were the same as those specified in the Common Test Conditions [6] of the 3DV Core Experiments conducted by (JCT-3V). All sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bit per sample.

The sequences were compressed with MVC+D and 3D-AVC using 3D-ATM v9.0 [7] under the conditions defined in [3, 6] (see Table 1). For each sequence, two stereo pair configurations were considered: a stereo pair formed from two synthesized views, referred to as SS in this paper, and a stereo pair formed from one original (decoded) view and one synthesized view, referred to as OS in this paper. For each codec, four rate points were considered. Additionally, a "reference" stereo pair was generated from the original data for each stereo pair configuration. Thus, this resulted in a set of $4 \times 2 \times (1 + 2 \times 4) = 72$ test stimuli. The synthesized views were generated using VSRS-1D-Fast v8.0 [8], under the conditions defined in [3, 6] (see Table 1).

Additionally, five training samples were generated using the *Poznan CarPark* sequence with similar conditions and manually selected by expert viewers so that the quality of samples were representative of all grades of the rating scale.

2.2. Test methodology

The single-stimulus (SS) methodology [9, 10] was chosen. A five-grade numerical categorical scale was used. The rating scale ranged from 1 to 5, with 1 indicating the lowest quality and 5 indicating the highest quality. After the presentation of each video sequence, a five-second voting time followed. Subjects were asked to rate the overall quality of the video sequence to be evaluated, and to express these judgments in terms of the wordings used to define the rating scale.

2.3. Test design

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. Since the total number of test samples was too large for a single test session, the overall experiment was split into two sessions of approximately ten minutes each. Between the sessions, the subjects took a ten minutes break. The test material was randomly distributed over the two test sessions.

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 each test session to stabilize the subjects' ratings. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each group of subjects, whereas the same content was never shown consecutively.

2.4. Test environment

In total, three laboratories conducted the subjective evaluation. All laboratories fulfill the recommendations for the subjective evaluation of visual data issued by ITU-R [9]: each test room is equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of the maximum screen luminance, whereas the color of all the background walls and curtains present in the test area are in mid grey.

Table 2: Test environment.

	Lab1	Lab2	Lab3
Monitor	Hyundai S465D	LG 47LM660	Sony KDL-55X9005A
Size	46"	47"	55"
#Subjects (♂/♀)	22 (15/7)	24 (14/10)	18 (16/2)
Age (average)	$20-31$ (23.1)	$18-28$ (19.6)	$20-31$ (26.5)

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 about 3 times the picture height. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively. The main differences between the laboratories were in terms of display characteristics and number of observers (see Table 2).

3. STATISTICAL ANALYSIS

The subjective results were processed by first detecting and removing subjects whose scores appeared to deviate strongly from others. The outlier detection was performed according to the guidelines described in Section 2.3.1 of Annex 2 of [9] and for each laboratory independently. In this study, one outlier was detected in Lab1 and one outlier was detected in Lab2. Then, for each laboratory, 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 [11], a regression was fitted to each $\left[MOS^{LabX}, MOS^{LabY}\right]$ data set using cubic fitting. 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

$$\left|\widehat{MOS}_i^{LabX} - MOS_i^{LabY}\right| > CI_i^{LabX} + CI_i^{LabY}$$

where CI_i^{LabX} and CI_i^{LabY} are the 95% CIs corresponding to $\widehat{MOS}_i^{LabX}$ and MOS_i^{LabY}, respectively, and $\widehat{MOS}_i^{LabX}$ is the fitted MOS.

3.2. Estimation errors

To determine whether the difference between two sets of scores corresponding to the same stereo pair evaluated in two different laboratories 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 [12], 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 [13], 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 two stereo pairs evaluated in the same laboratory is statistically significant, a multiple comparison test was performed similarly to Sec. 3.2.

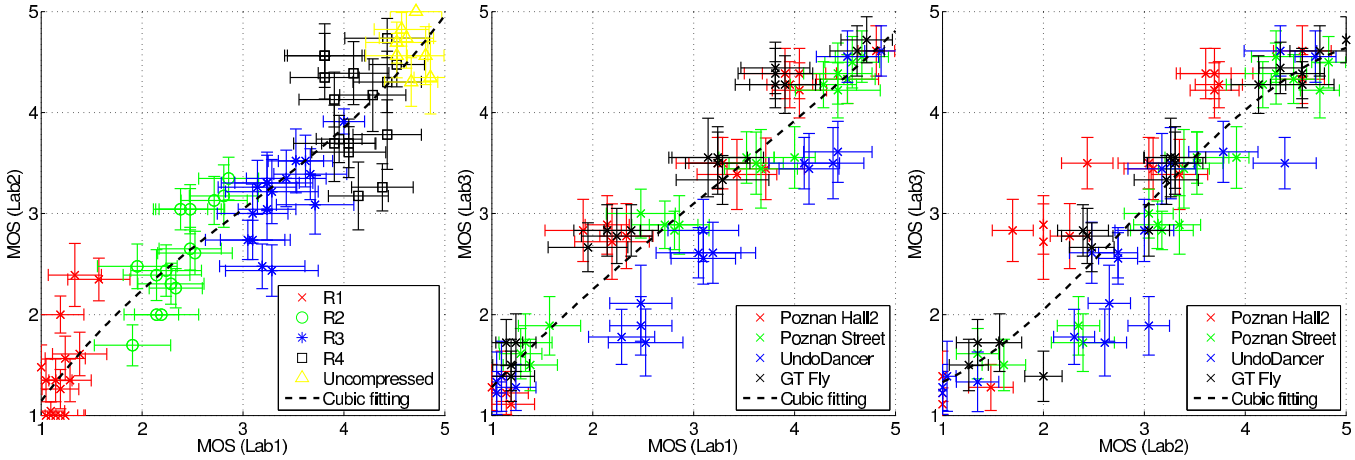
3.4. Average bit rate difference

The average bit rate difference for 3D-AVC over MVC+D was computed using the model proposed in [14]. This model is an extension of the Bjøntegaard model for subjective scores: ΔR is the average bit rate difference computed from the MOS; $[\Delta R_{\min}, \Delta R_{\max}]$ provide a confidence interval on ΔR and is determined considering the CI computed on the subjective scores; the confidence index takes into account the spreading of the MOS over the rating scale and the goodness of the fit of the values.

4. RESULTS AND DISCUSSION

Figure 1 shows the scatter plots comparing the results of the different laboratories. The horizontal and vertical error bars represent the CI corresponding to the laboratories on the x- and y-axis, respectively. The data points are colored based on the different contents or rate points for better visualization. The cubic regressions fitted to each data set are represented to illustrate the trend of the data points.

Ideally, all points would be on a $45°$ line if the MOS values for each condition were the same between two laboratories. However, some points lie above the line, whereas others lie below. For example, subjects in Lab3 graded content *UndoDancer* at rate points R2, R3, and R4 lower than subjects in Lab1. Similarly, subjects in Lab3 graded content *Poznan Hall2* at rate points R2, R3, and R4 lower than subjects in Lab2. Nevertheless, no significant systematic offset can be observed between the MOS values of the different laboratories, which means that, in overall, subjects did not score more



(a) Lab2 versus Lab1 (b) Lab3 versus Lab1 (c) Lab3 versus Lab2

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

Table 3: Performance indexes.

	PLCC			SROCC			RMSE			OR		
	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3
Lab1	–	0.9461	0.9429	–	0.9393	0.9340	–	0.3962	0.4073	–	20.83%	16.67%
Lab2	0.9407	–	0.9321	0.9399	–	0.9356	0.3911	–	0.4177	27.78%	–	19.44%
Lab3	0.9430	0.9294	–	0.9340	0.9356	–	0.3737	0.4146	–	12.50%	20.83%	–

Table 4: Estimation errors.

	Correct estimation			*Underestimation*			*Overestimation*		
	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3
Lab1	–	94.44%	97.22%	–	4.17%	2.78%	–	1.39%	0%
Lab2	93.06%	–	93.06%	2.78%	–	4.17%	4.17%	–	2.78%
Lab3	98.61%	88.89%	–	0%	6.94%	–	1.39%	4.17%	–

pessimistically nor more optimistically between the different laboratories.

Regarding the comparison between Lab1 and Lab3, the cubic fitting is close to a straight line, but its slope is smaller than $45°$. This indicates that subjects in Lab3 graded low quality stimuli higher than subjects in Lab1, whereas subjects in Lab1 graded high quality stimuli higher than subjects in Lab3. Regarding the comparison between Lab1 and Lab2 as well as between Lab2 and Lab3, the fitted cubic curves exhibit a sigmoid shape, which indicates non-linearity between the results of the different laboratories. For example, the ranges of grades associated with rate points R1 and R4 are wider in Lab2 than in Lab1. Nevertheless, the cubic regressions do not deviate much from a straight line.

As the mapping of MOS^{LabX} to MOS^{LabY} yields slightly different results when compared to mapping of MOS^{LabY} to MOS^{LabX}, both mappings are considered in the following subsections and results are reported for both cases. A value $v(i,j)$ on row i and column j is computed considering mapping of MOS^{Labi} to MOS^{Labj}.

4.1. Performance indexes

Table 3 reports the performance indexes. Results show that there is a strong correlation between the different laboratories, as the correlation indexes are above 0.92 in all cases. The PLCC, SROCC, and RMSE indexes are similar in all cases. However, the OR index shows a wider variation between the different cases. In particular, the OR values when mapping the results of Lab2 to Lab1 and Lab3 to Lab1 are above 27% and below 13%, respectively, whereas the average OR value is about 20% in the other cases. These results indicate that the correlation between Lab1 and Lab3 is the strongest.

4.2. Estimation errors

Table 4 reports the estimation errors. Results again show that there is a strong correlation between the different laboratories; especially between Lab1 and Lab3 (*Correct estimation* above 97%). However, when mapping the results of Lab3 to those of Lab2, the *Correct estimation* is below 89%, whereas the *Underestimation* and *Overestimation* are above 4%.

Table 5: Classification errors.

	Correct decision			*False ranking*			*False differentiation*			*False tie*		
	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3
Lab1	–	82.20%	79.50%	–	0%	0%	–	7.63%	7%	–	10.17%	13.81%
Lab2	80.99%	–	78.09%	0%	–	0.08%	7.86%	–	8.76%	11.15%	–	13.07%
Lab3	79.03%	78.48%	–	0%	0.20%	–	10.95%	10.45%	–	10.02%	10.88%	–

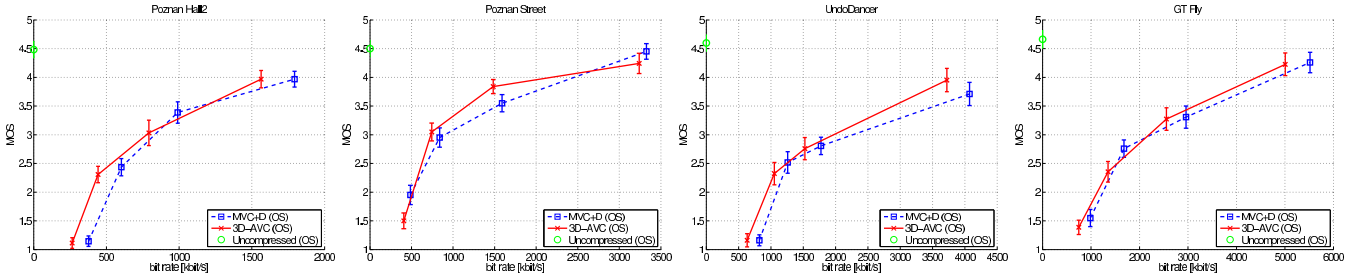


Fig. 2: Rate distortion curves: OS stereo pair.

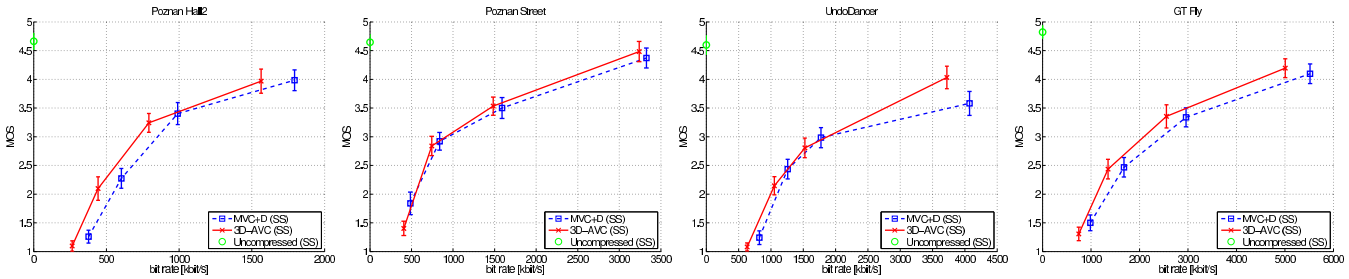


Fig. 3: Rate distortion curves: SS stereo pair.

4.3. Classification errors

Table 5 reports the classifications errors. About 80% of all possible distinct combinations of two stereo pairs lead to the same conclusion in different laboratories. Moreover, *False ranking*, which is the most offensive error, almost never occurs. *False tie* occurs in more than 10% of the cases, but this is the least offensive error. Results for *False differentiation* are in overall lower between Lab1 and Lab2 than between Lab1 and Lab3, which indicates that the correlation between Lab1 and Lab2 is higher than between Lab1 and Lab3, as opposed to the results of the performance indexes. However, the difference is not as big as for the performance indexes.

4.4. Rate distortion curves

The previous results show a strong correlation between the different laboratories. To further determine whether the scores from the different laboratories can be merged, an ANOVA was performed at a 5% significance level on the raw scores. The main effect of laboratories was not significant (see Table 6). Therefore, the raw scores from the three laboratories were merged in the following analyses.

Figures 2 and 3 depict the rate distortion curves for the SS and OS stereo pairs, respectively. As it can be observed, 3D-AVC usually outperforms MVC+D, as most of the R-D curve of 3D-AVC lie above that of MVC+D. However, comparing the two codecs at specific QP settings show that the confidence intervals overlap in most cases, which indicates that the difference between the two codecs is not significant in most cases.

4.5. Average bit rate difference

Since the resulting bit rates for the same QP are different between MVC+D and 3D-AVC, the model proposed in [14] provides a useful analysis of the performance of the two codecs. Table 7 reports the average bit rate difference ΔR for 3D-AVC over MVC+D computed from the MOS. For both stereo pair configurations, results show that, in average, 3D-AVC offers 14% bit rate reduction when compared to MVC+D,

Table 6: ANOVA: main effect of laboratories.

Laboratories	1 vs 2	1 vs 3	2 vs 3	1 vs 2 vs 3
p-value	0.52	0.13	0.33	0.31

Table 7: Bit rate differences for 3D-AVC over MVC+D.

Sequence	SS stereo pair			OS stereo pair		
	ΔR	$[\Delta R_{\min}, \Delta R_{\max}]$	Confidence index	ΔR	$[\Delta R_{\min}, \Delta R_{\max}]$	Confidence index
Poznan Hall2	-17%	$[-31\%, -1\%]$	89%	-13%	$[-29\%, 3\%]$	87%
Poznan Street	-8%	$[-25\%, 17\%]$	93%	-21%	$[-35\%, -3\%]$	84%
UndoDancer	-14%	$[-25\%, 10\%]$	88%	-15%	$[-32\%, 8\%]$	82%
GT Fly	-16%	$[-32\%, 5\%]$	89%	-6%	$[-25\%, 19\%]$	87%
Average	-14%	$[-28\%, 8\%]$	90%	-14%	$[-30\%, 7\%]$	85%

which is lower than the 22.6% bit rate reduction measured based on objective results [15]. However, from the confidence intervals, it can be seen that the bit rate difference varies from -30% to $+8\%$, which indicates that sometimes MVC+D is better than 3D-AVC, as it can observed from the rate distortion curves.

5. CONCLUSION

In this paper, the results of a 3D subjective assessment test conducted at three different test laboratories were analyzed, with the aim of verifying the consistency of the results and the impact (if any) of the laboratory set-up at each test site. The raw data coming from the three laboratories were submitted to a cross-laboratory analysis to determine if consistent results are obtained when executing a test applying different set-up, e.g., different display and different subjects. Statistical analyses, similar to those commonly applied when benchmarking objective video quality metrics, i.e., Pearson and Spearman correlation coefficients, root-mean-square error, and outlier ratio, were applied. Estimation and classification errors were also considered to estimate the differences when comparing subjective scores between pairs of laboratories. These analyses show that laboratories employing different displays and different subjects could still produce highly correlated results, if the test plan is well designed and the test are conducted following the same guidelines. In our experiment the three participating laboratories showed high correlation of the results of the subjective evaluations on the same content, regardless the model and the size of the 3D display and using different test subjects.

6. REFERENCES

[1] A. Perkis, J. You, L. Xing, T. Ebrahimi, F. De Simone, M. Rerabek, P. Nasiopoulos, Z. Mai, M. Pourazad, K. Brunnstrom, K. Wang, and B. Andren, "Towards certification of 3D video quality assessment," in *6th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, January 2012.

[2] 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*.

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

[4] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori, "Overview of the MVC+D 3D video coding standard," *Journal of Visual Communication and Image Representation*, vol. 25, no. 4, pp. 679–688, 2014.

[5] Ying Chen and Anthony Vetro, "Next-generation 3d formats with depth map support," *IEEE MultiMedia*, vol. 21, no. 2, pp. 90–94, April 2014.

[6] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, "Common Test Conditions of 3DV Core Experiments," Doc. JCT3V-E1100, Vienna, Austria, August 2013.

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

[8] 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.

[9] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, January 2012.

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

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

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

[13] 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)*, July 2013.

[14] P. Hanhart and T. Ebrahimi, "Calculation of average coding efficiency based on subjective quality scores," *Journal of Visual Communication and Image Representation*, vol. 25, no. 3, pp. 555–564, 2014, QoE in 2D/3D Video Systems.

[15] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, "AHG9: 3D-AVC Coding Results," Doc. JCT3V-F0094, Geneva, Switzerland, October 2013.