ON THE EVALUATION OF 3D CODECS ON MULTIVIEW AUTOSTEREOSCOPIC DISPLAY

Philippe Hanhart and Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

ABSTRACT

Evaluating the performance of different 3D codecs on multiview autostereoscopic monitor is a tedious process, as it requires the synthesis of a dense set of views. Therefore, it is legitimate to ask if evaluations could be performed on stereoscopic monitors instead and could lead to similar results as on multiview autostereoscopic monitors. This paper tries to answer to this question by analyzing a set of subjective scores resulting from evaluations of different 3D codecs on both display technologies. Results show that the comparison of different 3D codecs on stereoscopic display leads to similar conclusions when compared to multiview autostereoscopic display.

Index Terms— 3D, subjective quality assessment, multiview autostereoscopic display, stereoscopic display

1. INTRODUCTION

Multiview autostereoscopic monitors are expected to be among key elements in bringing 3D into the home. The Video Coding Experts Group (VCEG) and Moving Picture Experts Group (MPEG) have recently joined their efforts to develop new 3D video compression standards to enable both advanced stereoscopic display processing and improved support for multiview autostereoscopic displays [1]. However, to evaluate the performance of different 3D codecs on multiview autostereoscopic monitors, it is necessary to synthesize and interleave a dense set of views, which requires a lot of time, processing power, and storage capacity. Therefore, it is legitimate to ask if evaluations could be performed on stereoscopic monitors instead and could lead to similar results as on multiview autostereoscopic monitors.

In March 2011, a Call for Proposals (CfP) on 3D Video Coding Technology was issued by MPEG [2]. To support multiview autostereoscopic displays, a 3-view configuration was assumed. The decoded data, i.e., texture views and corresponding depth maps, was used to synthesize a set of virtual views at selected positions. The decoded and synthesized views were displayed on a multiview autostereoscopic monitor. The 3-view configuration was evaluated on both stereoscopic and multiview autostereoscopic displays. In the first case, the displayed stereo pair was formed from two synthesized views. In the second case, a dense set of 28 synthesized views was displayed on a multiview autostereoscopic monitor. Each compression algorithm was subjectively evaluated on both display technologies.

Stankiewicz and Wegner [3] have analyzed the subjective scores resulting from the evaluation of the responses to the MPEG CfP. They have shown that the mean opinion scores obtained on stereoscopic and multiview autostereoscopic displays were highly correlated according to Pearson and Spearman correlation coefficients (> 0.94). In this paper, we further analyze the subjective scores obtained on stereoscopic and multiview autostereoscopic monitors to determine whether there is an absolute or relative correspondence between the scores obtained on the two display technologies.

The remainder of the paper is organized as follows. The methodology followed to analyze the subjective scores is described in Section 2. Results are reported and analyzed in Section 3. Finally, concluding remarks are given in Section 4.

2. METHODOLOGY

In this paper, mean opinion scores (MOS) and corresponding 95% confidence intervals (CI) that were computed by the MPEG test coordinator on a total of 36 naïve viewers from three different laboratories [4] have been used. Outlier detection was performed by the MPEG test coordinator. As the number of valid subjects for each condition is not specified, we assumed a total of 36 valid subjects. We further assumed that the MOS and CI values were computed according to recommendation ITU-R BT.500-13 [5].

2.1. Estimation errors

To determine whether the difference between two MOS corresponding to the same decoded 3D data evaluated on stereoscopic and multiview autostereoscopic monitors is statistically significant, a two-sample unpooled t-test was performed as the score distributions have unknown and unequal variances. If the observed value was inside the critical region determined by the 95% two-tailed Student's t-distribution, then the two MOS values were considered to be statistically different at a 5% significance level. The percentage of *Correct Estimation*, *Underestimation*, and *Overestimation* were recorded from all possible combinations of content, codec, and bit rate.

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.

2.2. Classification errors

In recommendation ITU-T J.149 [6], it is suggested to compute the classification errors to evaluate the performance of an objective metric. A classification error is made when the objective metric and subjective test lead to different conclusions on a pair of video sequences, *A* and *B*, for example. In this paper, this methodology is extended to the case of comparison of a pair of subjective tests, *A* and *B*, corresponding to quality assessment of 3D content on a stereoscopic and a multiview autostereoscopic monitor. Three types of error can happen:

a) *False Tie*, the least offensive error, which occurs when the evaluation on multiview autostereoscopic monitor says that *A* and *B* are different whereas the evaluation on stereoscopic monitor says that they are identical,
b) *False Differentiation*, which occurs when the evaluation on multiview autostereoscopic monitor says that *A* and *B* are identical whereas the evaluation on stereoscopic monitor says that they are different,
c) *False Ranking*, the most offensive error, which occurs when the evaluation on multiview autostereoscopic monitor says that *A* (*B*) is better than *B* (*A*) whereas the evaluation on stereoscopic monitor says the opposite.

To determine whether the difference between two MOS corresponding to a pair of decoded 3D data evaluated on the same display technology is statistically significant, a two-sample unpooled *t*-test was performed similarly to Section 2.1. The percentage of *Correct Decision*, *False Tie*, *False Differentiation*, and *False Ranking* were recorded from all possible distinct pairs of decoded 3D data, i.e., combination of content, codec, and bit rate.

3. RESULTS

Table 1 gives the estimation errors for class A (1024 × 768 pixels, 30 fps) and class C (1920 × 1088 pixels, 25 fps) contents separately, as well as for all contents together. In average, only about 40% of all possible combinations of content, codec, and bit rate had statistically equivalent MOS on stereoscopic and multiview autostereoscopic monitors, whereas the MOS were either underestimated or overestimated on the stereoscopic monitor in about 60% of the cases. In particular, for class C, about half of the decoded 3D data was underestimated on the stereoscopic monitor when compared to the multiview autostereoscopic monitor. Therefore, we conclude that there is no absolute correspondence between the scores obtained on the two display technologies.

Table 1. Estimation errors.

	Correct Estimation	*Overestimation*	*Underestimation*
Class A	42.19%	25.26%	32.55%
Class C	37.76%	12.76%	49.48%
All	39.97%	19.01%	41.02%

Table 2. Classification errors.

	Correct Decision	*False Ranking*	*False Differentiation*	*False Tie*
Class A	82.82%	3.45%	6.52%	7.21%
Class C	84.36%	3.04%	6.60%	6.00%
All	83.13%	3.51%	6.68%	6.68%

Table 2 gives the classification errors for class A and class C contents separately, as well as for all contents together. On all contents, around 83% of all possible distinct pairs of decoded 3D data lead to the same conclusion on stereoscopic monitor when compared to multiview autostereoscopic monitor. *False Ranking* occurs in only 3.5% of the cases. The classification errors are relatively similar across class A, class C, and all contents. Therefore, we conclude that there is a relative correspondence between the scores obtained on the two display technologies. These results show that the comparison of different 3D codecs on stereoscopic monitor leads to similar results when compared to comparison on multiview autostereoscopic monitor.

4. CONCLUSION

In this paper, we investigated the estimation and classification errors resulting from subjective evaluation of 3D codecs on a stereoscopic monitor instead of a multiview autostereoscopic monitor. The stereo pairs were formed from two synthesized views, whereas a dense set of 28 views was displayed on the multiview autostereoscopic monitor. Results show that there is a relative correspondence between the scores obtained on the two display technologies, whereas there is no absolute correspondence. These results indicate that the comparison of different 3D codecs on stereoscopic monitor leads to similar conclusions when compared to multiview autostereoscopic monitor. Therefore, we suggest to evaluate the performance of 3D codecs on stereoscopic monitors, as the generation of a dense set of views requires a lot of time, processing power, and storage capacity.

5. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Applications and Requirements on 3D Video Coding," Doc. N12035, Geneva, CH, Mar. 2011.

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

[3] T.G.O. Stankiewicz and K. Wegner, "Correlation analysis between MOS data collected on stereoscopic and autostereoscopic displays," Tech. Rep. JCT3V-C0202, JCT-3V, Geneva, CH, Jan. 2013.

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

[5] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," ITU, Jan. 2012.

[6] ITU-T J.149, "Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)," ITU, Mar. 2004.