

**Title:** AHG9: On the evaluation on stereoscopic and multiview autostereoscopic displays

**Status:** Input Document

**Purpose:** Proposal

**Author(s) or Contact(s):** Philippe Hanhart  
Touradj Ebrahimi

**Email:** philippe.hanhart@epfl.ch

**Source:** Ecole Polytechnique Fédérale de Lausanne (EPFL), COST Action IC1003 - Qualinet

## Abstract

In this contribution, the estimation and classification errors resulting from subjective evaluation on a stereoscopic monitor (with passive polarized glasses) instead of a multiview autostereoscopic monitor are reported. A set of subjective data, which was collected during the formal evaluation of the 3DVC proposals on the 3-view configuration, is used as ground truth. It is reported that there is a relative correspondence between the scores obtained on the two display technologies, and that the comparison of different 3D codecs on stereoscopic display lead to similar results as on multiview autostereoscopic display. Therefore, it is suggested to evaluate the codec performance on stereoscopic display only, as conducting the evaluations on both display technologies is very time and effort consuming.

## 1 Introduction

In the 3-view configuration, as considered in the 3DVC Call for Proposals (CfP) (N12036), three cameras are used to produce the input views at the encoder side. 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, as specified in Table 1. In the latter case, a dense set of 28 synthesized views was displayed on the multiview autostereoscopic monitor, as specified in Table 1. Therefore, each compression algorithm was subjectively evaluated on both display technologies. Each time, mean opinion scores (MOS) and associated 95% confidence intervals (CI) were computed from the individual scores given by a total of 36 subjects (N12347).

Table 1: Synthesized output views for stereoscopic and autostereoscopic monitors.

Seq. ID	Test Sequence	Test Class	Input views	Stereo pair	Views for autostereoscopic display
S01	Poznan_Hall2	A	7-6-5	6.125-5.875	All 1/16 positions between views 7 and 5
S02	Poznan_Street		5-4-3	4.125-3.875	All 1/16 positions between views 5 and 3
S03	Undo_Dancer		1-5-9	4.5-5.5	All 1/4 positions between views 1 and 9
S04	GT_Fly		9-5-1	5.5-4.5	All 1/4 positions between views 9 and 1
S05	Kendo	C	1-3-5	2.75-3.25	All 1/8 positions between views 1 and 5
S06	Balloons		1-3-5	2.75-3.25	All 1/8 positions between views 1 and 5
S07	Lovebird1		4-6-8	5.75-6.25	All 1/12 positions between views 4 and 8
S08	Newspaper		2-4-6	3.75-4.25	All 1/12 positions between views 2 and 6

To evaluate the performance of different 3D codecs on multiview autostereoscopic monitor, it is necessary to synthesize and interleave a dense set of views, which requires a lot of time, processing power, and storage capacity. Moreover, conducting the evaluations on both stereoscopic and multiview autostereoscopic monitors is very time consuming and expensive. Therefore, it is legitimate to ask if evaluations could be performed on stereoscopic monitor only and could lead to similar results as on multiview autostereoscopic monitor. It was reported in JCT3V-C0202 that the MOS obtained on stereoscopic and multiview autostereoscopic displays were highly correlated in terms of the Pearson and Spearman correlation coefficients. In this contribution, the subjective scores obtained on stereoscopic and

multiview autostereoscopic monitors are further analyzed to determine whether there is an absolute or relative correspondence between the scores obtained on the two display technologies. It is reported that the MOS obtained on stereoscopic and multiview autostereoscopic monitors for the same decoded 3D data, i.e., texture views and associated depth maps, are statistically identical in only 40% of the cases. Therefore, it is concluded that there is no absolute correspondence between the scores obtained on the two display technologies. However, when comparing a pair of decoded 3D data to determine whether the perceived quality is worse, equal, or better, it is reported that the evaluations on stereoscopic and multiview autostereoscopic monitors would lead to the same conclusion in 83% of the cases. Therefore, it is concluded that there is a relative correspondence between the scores obtained on the two display technologies, and that the comparison of different 3D codecs on stereoscopic monitor lead to similar results when compared to comparison on multiview autostereoscopic monitor.

## 2 Methodology

In this contribution, MOS and CI values that were computed by the MPEG test coordinator on a total of 36 naïve viewers from three different laboratories (N12347) were used. 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. As the number of valid subjects for each condition is not specified, a total of  $n = 36$  valid subjects were assumed. It was further assumed that the MOS and CI values were computed according to recommendation ITU-R BT.500-13, where the MOS and 95% CI are defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$[\bar{x} - \delta, \bar{x} + \delta]$$

where

$$\delta = 1.96 \frac{s}{\sqrt{n}}$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\bar{x} - x_i)^2}$$

### 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.

The observed value  $t_{obs}$  was computed from the observations for each comparison

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where  $\bar{x}_1$  and  $\bar{x}_2$  are the two MOS corresponding to the stereoscopic and multiview autostereoscopic monitors, respectively,  $s_1$  and  $s_2$  are the corresponding sample standard deviation, and  $n_1 = n_2 = 36$ .

If the observed value  $t_{obs}$  was inside the critical region determined by the 95% two-tailed Student's  $t$ -distribution with  $df$  degrees of freedom, then the two MOS values were considered to be statistically different at a 5% significance level.

$$df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left( \frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left( \frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

Table 2: Interpretation of the statistical test.

Observed value	Conclusion	Result
$t_{obs} > t\left(1 - \frac{\alpha}{2}, df\right)$	$\bar{x}_1 > \bar{x}_2$	Overestimation
$t\left(\frac{\alpha}{2}, df\right) \leq t_{obs} \leq t\left(1 - \frac{\alpha}{2}, df\right)$	$\bar{x}_1 = \bar{x}_2$	Correct Estimation
$t_{obs} < t\left(\frac{\alpha}{2}, df\right)$	$\bar{x}_1 < \bar{x}_2$	Underestimation

The percentage of Correct Estimation, Underestimation, and Overestimation were recorded from all possible combinations of content, codec, and bit rate.

## 2.2 Classification errors

In recommendation ITU-T J.149, 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 contribution, 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:

- 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,
- 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,
- 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.

Table 3: Classification errors.

		Autostereo		
		$MOS_A > MOS_B$	$MOS_A = MOS_B$	$MOS_A < MOS_B$
Stereo	$MOS_A > MOS_B$	Correct Decision	False Differentiation	False Ranking
	$MOS_A = MOS_B$	False Tie	Correct Decision	False Tie
	$MOS_A < MOS_B$	False Ranking	False Differentiation	Correct Decision

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 4 gives the estimation errors for class A and class C 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 under estimated 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, it is concluded that there is no absolute correspondence between the scores obtained on the two display technologies.

Table 4: Estimation errors.

	<b>Correct Estimation</b>	<b>Overestimation</b>	<b>Underestimation</b>
<b>Class A</b>	42.19%	25.26%	32.55%
<b>Class C</b>	37.76%	12.76%	49.48%
<b>All</b>	39.97%	19.01%	41.02%

Table 5 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 as when compared to multiview autostereoscopic monitor. False Ranking occurred in only 3.5% of the cases. The classification errors are relatively similar across class A, class C, and all contents. Therefore, it is concluded that there is a relative correspondence between the scores obtained on the two display technologies, and that the comparison of different 3D codecs on stereoscopic monitor leads to similar results when compared to comparison on multiview autostereoscopic monitor.

Table 5: Classification errors.

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

### 4 Conclusion

In this contribution, the estimation and classification errors resulting from subjective evaluation on a stereoscopic monitor instead of a multiview autostereoscopic monitor were investigated. It is reported that there is a relative correspondence between the scores obtained on the two display technologies, and that the comparison of different 3D codecs on stereoscopic monitor leads to similar results as on multiview autostereoscopic monitor. Therefore, it is suggested to evaluate the codec performance on stereoscopic display only, as conducting the evaluations on both display technologies is very time and effort consuming.

### 5 Acknowledgement

This work has been conducted in the framework of the Swiss National Science Foundation (grant 200021-143696-1) and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET.

### 6 References

ISO/IEC JTC1/SC29/WG11, "Call Proposals on 3D Video Compression Technology," Doc. N12036, Geneva, Switzerland, March 2011.

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.

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

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

JCT-3V, "Correlation analysis between MOS data collected on stereoscopic and autostereoscopic displays," JCT3V-C0202, JCT-3V Meeting, Geneva, Switzerland, January 2013.

## **7 Patent rights declaration(s)**

**EPFL does not have any current or pending patent rights relating to the technology described in this contribution.**