



Exemplary Test Design and Evaluation of an Autostereoscopic 3DTV considering Display Operating Parameters

Ruth Schultheis¹, Sara Kepplinger², Frank Hofmeyer¹, Nikolaus Hotton³

¹ Technische Universität Ilmenau, Germany

² Fraunhofer Institute of Digital Media Technology IDMT, Germany

³ Hochschule Furtwangen University, Germany

¹ruth.schultheis@tu-ilmenau.de, ²sara.kepplinger@idmt.fraunhofer.de, ³hon@hs-furtwangen.de

Abstract

This paper addresses the influences of standard operating parameters like backlight, contrast, sharpness, etc. of an autostereoscopic display on subjective quality, in the absence of transmission quality impairments (best case reference). Therefore, a useful test environment is specified and applied. Results show that the influences of different display setups are distinguishable from each other but, test sequences have a crucial impact on this. In a further step other configurations and displays should be investigated.

Index Terms: quality evaluation; autostereoscopic display; video impression; impact of display operating parameters

1. Introduction

The question about how to define the optimal test setup for subjective quality evaluation of 3D representations, in the context of Quality of Experience (QoE), which is defined in [1], is not new (see, e.g. [2], [3], [4] – [10]). This includes consideration of test environment, contextual influences, methodological approaches in general (including different ways of scaling and level definition), useful quality attributes in particular, as well as the choice of adequate test stimuli. Some results found consideration in standards partly and lately (see new Recommendations [11] and [12]). However, remaining questions include the amount of influence by the used display and its operating parameters. Current standards of 3D video quality assessments often refer to coding and transmission error scenarios where the 3D video sequences are moderately to strongly degraded [11], [13]. Due to our research in quality assessment of autostereoscopic representations, we are always in search of test methods which are significant and improved under best case conditions to determine the general impact of display operating parameters on observers' quality of experience. Towards this, several subjective evaluations were performed under best case payout conditions with isolated but well described display operating parameters to check their specific impact on overall viewing experience. In the long run, the goal is to identify the general effects of different kinds of 3D display technology and their operating parameters on variable user groups (e.g. naïve or experienced) under distinct viewing conditions (e.g. living room, exhibition hall). The current work focuses on these issues considering autostereoscopic display technology (AS3DTV) which is well described in [14] including a description of parallax barriers. We refer to this and to an overview provided by Woods [15], concerning the quality aspects of crosstalk. Furthermore, we are

looking for general test items suitable for such autostereoscopic assessments. In this paper we firstly focus on the influences of display operating parameters of a specific autostereoscopic multi view display and a suitable evaluation setup. The insights gained from this research also serve as basis of a later subjective quality assessment of autostereoscopic videos including gaze tracking data from observers.

2. Evaluation Setup and Design

The investigated research question is: "How strongly do different display operating parameters of an autostereoscopic display affect observers' Quality of Experience (QoE) under best case payout conditions?" To classify the outcome of this evaluation, the categories *overall quality*, *3d impression*, *disturbance by double images / crosstalk*, *perceived brightness* and *sharpness* are investigated. Here, the subjective perceived video quality is evaluated using different test sequences in distinct display operating configurations.

2.1. Test display and operating parameters

In this first study the performance of a *Tridality 55" ML5520va* [16] autostereoscopic display was evaluated. Its 3D technology is based on a parallax barrier. The display supports five multi views based on an overall full High Definition (HD) resolution of 1920x1080 pixels. The view patterns are V3, V4, V5, V1, V2, V3, V4, [16]. The resolutions of the test stimuli used are represented in table 3.

2.1.1. Tested display configurations

The adjustable operating parameters of the evaluated autostereoscopic display were a) backlight, b) contrast, c) sharpness, d) brightness, e) color shade and f) color. Each parameter could be adjusted on a scale between 0 and 100 without any explicit unit. However, table 1 shows measureable values (measured brightness [cd/m²]) for the parameters backlight and brightness in order to get a feel of the provided scales by the display. For these measurements a white and grey test chart, a HCFR colorimeter (software) with an X-Rite i1 display colorimeter (hardware) as well as a MINOLTA luminance meter LS-100, was used.

The evaluated display was able to save configurations in three pre-sets. Each pre-set was additionally fixed to a color temperature of 10000 K, a color control of 0, a native gamma correction and a negative adaptive contrast. The aspect ratio was adjusted permanently to 1:1 related to its overall spatial resolution of 1920x1080 pixels. The volume was completely muted as no audio configuration was taken into account. Based

on several pre-tests looking at a whole range of possible combinations, six were identified as adequate with respect to QoE and chosen for evaluation (Table 2).

Tab. 1. *Brightness related to operating parameters a) and d)*

Test chart: 100% (white)					
Backlight (a)	100%	100%	100%	100%	100%
Brightness (d)	0%	25%	50%	75%	100%
[cd/m ²]	11.3	39	88	125	125
Test chart: 100% (white)					
Backlight (a)	50%	50%	50%	50%	50%
Brightness (d)	0%	25%	50%	75%	100%
[cd/m ²]	5.9	20	46	65	65
Test chart: 50% (grey)					
Backlight (a)	100%	100%	100%	100%	100%
Brightness (d)	0%	25%	50%	75%	100%
[cd/m ²]	0	4.2	23.6	62.3	121.5
Test chart: 50% (grey)					
Backlight (a)	50%	50%	50%	50%	50%
Brightness (d)	0%	25%	50%	75%	100%
[cd/m ²]	0	2.2	12.3	32.5	64


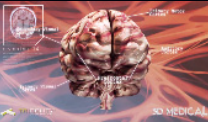





Tab. 2. *Display configurations used for the evaluation*

Configuration	Standard	Sharpness 0	Sharpness 100	Brightness 25	Brightness 75
Backlight	50	50	50	50	50
Contrast	50	50	50	50	50
Sharpness	50	0	100	50	50
Brightness	50	50	50	25	75
Color shade	50	50	50	50	50
Color	50	50	50	50	50

2.2. Test environment

The test environment usually refers to the viewing conditions recommended in ITU-T P.916 [11] and ITU-T P.910 [17]. Thus, a room illuminated using only artificial light, was chosen (see Figure 1). The display was attached to a stand, whose base was in line with the wall. The centre of the display was vertically positioned 1.5m from the ground. Two flood lights were placed on either side to ensure a background illumination of approximately 20 lux (as suggested in [17]). This background room illumination was frequently checked with a Minolta Illuminance Meter T-10. The viewing distance was set up to 3.9m, as recommended by Trideltity. The viewing position

Tab. 3: Spatial resolution (origin), resolution per view, frame rate and duration of test items. Last column indicates predicted subjective depth impression.

A) 	B) 	C) 	D) 
1080 x 2400 pixel, 960 x 960 pixel, 25 fps, 14 s	2048 x 1440 pixel, 1024 x 576 pixel, 30 fps, 18 s	2048 x 1440 pixel, 1024 x 576 pixel, 25 fps, 29 s	1936 x 1360 pixel, 968 x 544 pixel, 30 fps, 14s
E) 	F) 	G) 	A) middle B) middle C) middle – strong D) middle – strong E) slightly F) slightly – middle G) middle
2048 x 1440 pixel, 1024 x 576 pixel, 30 fps, 14 s	1936 x 1360 pixel, 986 x 544 pixel, 25 fps, 15s	2560 x 1800 pixel, 1280 x 720 pixel, 30 fps, 14 s	

(sweet spot) was thoroughly determined considering viewing angles and display position. It was perpendicularly measured from the display centre. A mark was set on the floor in order to constantly place a platform with a chair on it.

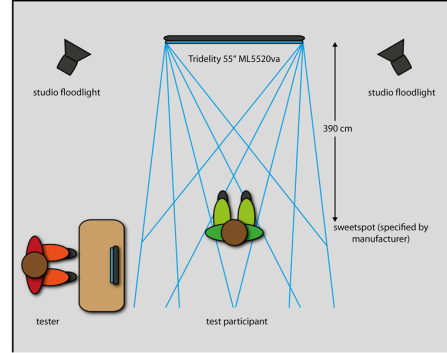


Figure 1: *Test environment and setup*

A LENOVO ideapad 305-15IBD laptop, based on an Intel CORE i5 5200U (10.0.10586.0) processor, including a solid state disk and an AMD Radeon R5 M330 (15.300.1025.1001) graphic card using a HDMI 1.4 interface was used as payout equipment. Windows 10 Home 64 Bit and AMD Radeon Crimson 15.11 were installed. The video setup of the laptop was fixed to a refresh rate of 60 Hz, a resolution of 1920x1080 pixels and a color depth of 32 bits with a basic color setup using brightness = 0, contrast = 50 and gamma = 1.0 for each color channel. The items were payout by the dedicated TriView 3D Player version 5.4.1.0.

2.3. Test stimuli

The items used in this first evaluation were specifically designed by Trideltity. The sequences are based on computer-generated imagery (CGI) animations and do not correspond to any stereoscopic video standards. Known technical parameters were solely spatial resolution, frame rate and duration. Unfortunately, induced absolute and relative parallaxes were completely unknown, so a predicted subjective depth impression had to be defined for each item based on a subjective off-testing by the test conductor (see Table 3).

2.4. Test method

To measure the perceived quality of each configuration, the test method Absolute Category Rating (ACR) was used to evaluate defined quality attributes. The five quality attributes were specified as follows:

- Overall quality: First impression of the video
- 3D impression: Stereoscopic perception of objects/scene
- Disturbance by double vision (crosstalk)
- Brightness: Visibility of the video relating to over- or underexposure (brightness may sometimes be a negative attribute, if the image is too bright to be recognized)
- Sharpness: Not pixelated video with smooth outlines of received objects

The voting by pen and paper was based on a quasi-continuous scale as described in Annex B of Recommendation ITU-T P.910 [17].

Attributes were scaled in 25 boxes ranging from bad to excellent, or, according to the category crosstalk, from very annoying to not annoying (only if crosstalk was indicated by observers). For analysis, those 25 boxes could be consolidated to five levels (5 excellent: boxes 21-25, 4 good: boxes 16-20, 3 fair: boxes 11-15, 2 poor: boxes 6-10, 1 bad: boxes 1-5) which would allow the definition of the mean opinion score (MOS) as suggested in [12].

2.5. Evaluation

The evaluation took place in February 2016 at the Technische Universität Ilmenau, and it was conducted in three phases: the preparation phase including a pre-test, the test session itself including a short break, and the closing phase. Overall, each test session was 30 to 40 minutes long. At the beginning the observers were tested regarding their stereoscopic and colour perception abilities using Randot stereotest [18] and Ishihara test [19]. Additionally demographic data of the observers were recorded. During the pre-test the test participants observed all seven test sequences in the *Standard* configuration. This was useful in order to familiarize observers with the test stimuli, the evaluation method, and to find their best fitting viewing position. After that, a first Simulator Sickness Questionnaire (SSQ) [20] was performed.

Throughout the test session all test participants were asked to evaluate the display quality subjectively, excluding possible influences by the image content of the test sequences. The test session was divided into two parts. In the first part all stimuli were randomly presented over three randomized display configurations. Those configurations were saved by the display provided pre-sets, which were randomized once more for every test sequence. Test participants were advised to watch the test stimuli unrestricted until they were ready for rating. After that the same test sequence was presented in a different display configuration. After a short break the second part of the test session was carried out in the same randomized manner presenting the remaining display configurations. In total the participants rated 42 combinations of test items versus display configuration. The test session ended with the second part of the SSQ.

3. Results

In total 21 participants (10 male, 11 female) between 22 and 48 years old were recruited for the test. The subjects were mostly

students of Technische Universität Ilmenau without any knowledge about the topic or the presented 3D display technology. All participants passed the Randot stereotest and Ishihara test, only one person showed a slight red-green colour deficiency. During assessment, eight subjects wore glasses, two subjects wore contact lenses.

The results (Figures 1 – 5) show that the display configurations Standard and Backlight were rated as good for the criteria overall quality, 3D impression and sharpness. Even so it seems that the configuration Backlight had a better rating for brightness than the configuration Standard. However Standard provoked less perceptible but not annoying crosstalk than the configuration Backlight. Similar results can be stated for the adjustment Sharpness 100. The increase of the sharpness value to 100 did not lead to a very different rating of the overall quality, 3D impression, sharpness or brightness. Crosstalk was indicated as slightly annoying (values between 6.5 and 16.5). The results according to the adjustments Brightness 25 and 75 represent that these configurations had a significant influence on the perception of the quality of the display. This is similar to the QoE of 2D displays. Both, 3D impression and sharpness received fair ratings regarding the attribute overall quality. They also revealed slightly annoying crosstalk values starting from 10 to 21 for Brightness 25 and to 19 for Brightness 75. Sharpness 0 had a big leverage on the delivered votes of the five quality factors. The configuration distinguished a fair overall quality with a median at 11 and brightness impression with a median achievement of 15. Furthermore, the 3D impression and sharpness were poor (median at 10 and 8). Most of the test participants agreed that this configuration caused “annoying double images” (values between 4 and 11.5). Overall, one can see the influences based on different display configurations. A significance test (Friedman), for the characteristic overall quality, with a presumption of 95 %, showed, that four out of six display configurations were influenced by the content of the test sequences, including Standard (p-value: 0.001), Brightness 25 (p-value: 0.021), Sharpness 0 (p-value: < 0.0001) and Sharpness 100 (p-value: 0.002). The results of the SSQ show that, at the beginning of the test, most of the participants were in a good physical condition. At the end some subjects felt slightly lightheaded but unfatigued. A few of the participants also expressed discontent over slight headache, difficulties in concentrating and problems with their vision, e.g. blurry view or acuteness of vision.

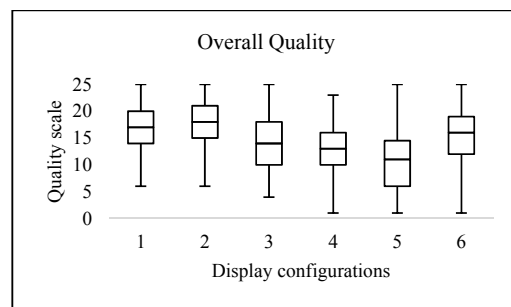


Figure 2: Evaluations of the overall quality. 1. Standard 2. Backlight 3. Brightness 25 4. Brightness 75 5. Sharpness 0 6. Sharpness 100

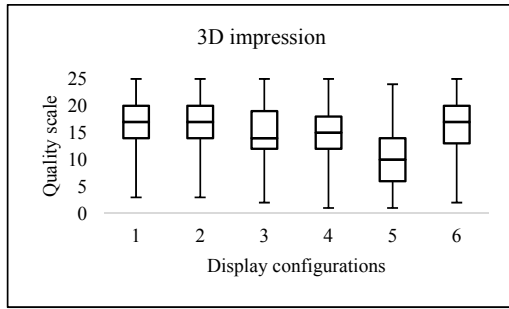


Figure 3: Evaluations of the 3D impression. 1. Standard 2. Backlight 3. Brightness 25 4. Brightness 75 5. Sharpness 0 6. Sharpness 100

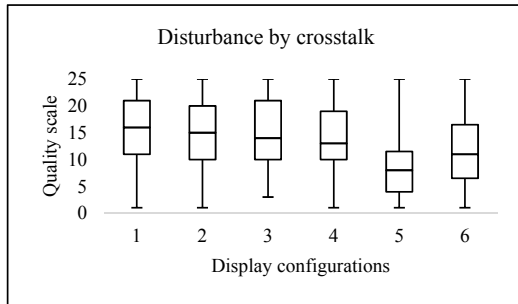


Figure 4: Evaluations of the disturbance of crosstalk. 1. Standard 2. Backlight 3. Brightness 25 4. Brightness 75 5. Sharpness 0 6. Sharpness 100

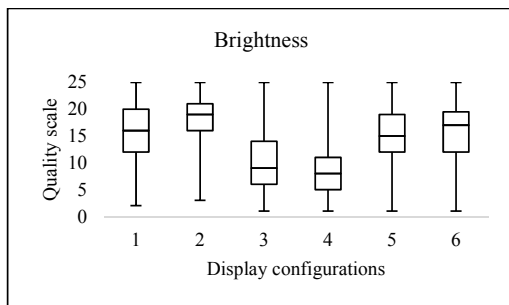


Figure 5: Evaluations of the brightness. 1. Standard 2. Backlight 3. Brightness 25 4. Brightness 75 5. Sharpness 0 6. Sharpness 100

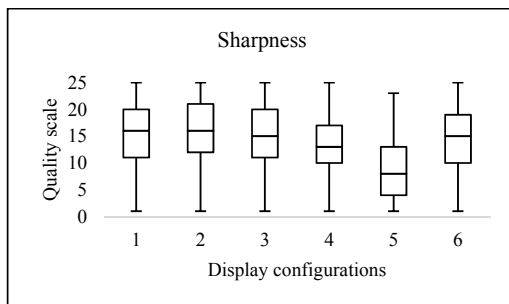


Figure 6: Evaluations of the sharpness. 1. Standard 2. Backlight 3. Brightness 25 4. Brightness 75 5. Sharpness 0 6. Sharpness 100

4. Discussion

The presented results show a test design and its outcome for the tested hypothesis about the impact of different display configurations of the AS3D display on subjective quality. The presented evaluation results lead to the assumption, that for AS3D as well the influence of brightness on quality perception is crucial. This confirms the results of previous S3DTV comparisons, e.g. [1]. The significance of the rating differences of chosen display configurations should be investigated further. As well as the interaction of the investigated characteristics and the scales. However, we could already detect useful results in order to prove investigated hypothesis and the test design. Regardless, it is questionable if and how much the image content of the test sequences had an impact on the subjects' assessments and this has to be investigated further especially as soon as more useful and well described test items are available. In addition it would be interesting to see the outcome of stereoscopic real shots, since in this test design only CGIs were used. Unfortunately, the presented evaluation was carried out in February 2016 before the publication of the new ITU-T Recommendation P.915 [11] in March 2016 which includes the information in section 7.3.1 on not increasing the number of levels as the resulting MOS does not improve.

5. Conclusions and Outlook

The purpose of this test was on the one hand, to evaluate the influence of different display operating parameters on subjective quality evaluation of an AS3D, and on the other hand, to test the evaluation design of a best-case payout scenario. In a first step, six display configurations of a specific display were chosen and evaluated. Test participants were asked to evaluate the test sequences with respect to five quality attributes. The test setup was defined based on existing guidelines and ITU Recommendations properly referring to coding and transmission error scenarios. However, there is no best-case Recommendation available addressing AS3D yet. Generally speaking, the presented test design might be useful when excluding the technological advantage of the AS3D display, namely, the possibility to change the viewers' position (instead of having only one sweet spot in the centre) or to allow multi viewing. Hence, a further adequate test design taking position changes into account might be advantageous. Furthermore, another issue is the availability of adequate test stimuli. One result is that content seems to be a relevant influencing factor on quality rating. Hence, in order to investigate the impacts of content and image characteristics beyond CGI, further work towards test item definition and creation as well as evaluation scenarios is necessary. This includes the well documented technical characteristics of the test item and its depth budget. This would allow the observation of the correlation between different test item characteristics with respect to quality rating, as well as the definition of thresholds.

6. Acknowledgements

We would like to thank all test participants and TRIDELITY AG for the videos we were allowed to use as test stimuli.

7. References

- [1] Qualinet White Paper on Definitions of Quality of Experience, 1st ed., European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), June 2012, output from the Dagstuhl seminar 12181.
- [2] Kepplinger, S., Hottong, N., Quality evaluation of stereo 3DTV systems with open profiling of quality. In: Proceedings of the 2014 IS&T/SPIE Conference on Electronic Imaging, Number 9014-46. IEEE Explore, San Francisco, CA, USA, 2014.
- [3] Chen, W., Fournier, J., Barkowsky, M. and Le Callet, P., "New requirements of subjective video quality assessment methodologies for 3DTV", Proc. 5th Intern. Wshp. on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2010), 2010.
- [4] Kepplinger, S., Jauch, C., Tobian, D., Defining the Viewing Conditions in the Home Environment and its Influences. In: Proc. of 4th Intern. Workshop on Perceptual Quality of Systems (PQS 2013), 2nd-4th September 2013, Vienna, Austria.
- [5] L. Xing, J. Xu, K. Skildheim, A. Perkis, and T. Ebrahimi, "Subjective Crosstalk Assessment Methodology for Auto-stereoscopic Displays," in 2012 IEEE International Conference on Multimedia and Expo, pp. 515–520.
- [6] Kaptein, R., Kuijsters, A., Lambooi, M., IJsselsteijn, W. A., & Heynderickx, I. (2008). Performance evaluation of 3D-TV systems.
- [7] Kulyk, V., Tavakoli, S., Folkesson, M., Brunnström, K., Wang, K., & Garcia, N. (2013). 3D Video Quality Assessment with Multi-scale Subjective Method.
- [8] Li, J. (2013). Methods for assessment and prediction of QoE, preference and visual discomfort in multimedia application with focus on S-3DTV. ED 503-211, University of Nantes, France.
- [9] Wang, K., Barkowsky, M., Brunnström, K., Sjöström, M., Cousseau, R., & LeCallet, P. (2012). Perceived 3D TV transmission quality assessment: Multi-laboratory results using Absolute Category Rating on Quality of Experience scale. IEEE Transactions on Broadcasting, 58(4), 544-557 (510.1109/TBC.2012.2191031).
- [10] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," EURASIP J Image Video Process, vol. 2013, no. 1, p. 50, 2013.
- [11] ITU-T Rec. P.914 (03/2016), Display Requirements for 3D Video Quality Assessment. Retrieved from ITU.
- [12] ITU-T Rec. P.915 (03/2016), Subjective assessment methods for 3D video quality. 2016, Retrieved from ITU.
- [13] ITU-R Rec. BT.2021 (02/2015), Subjective methods for the assessment of stereoscopic 3DTV systems. Retrieved from ITU.
- [14] H. Urey, K. V. Chellappan, E. Erden, and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," Proc. IEEE, vol. 99, no. 4, pp. 540–555, 2011.
- [15] A. J. Woods, "How are crosstalk and ghosting defined in the stereoscopic literature?" in Proceedings of SPIE Stereoscopic Displays and Applications XXII, pp. 78630Z-1 - 78630Z-12.
- [16] TRIDELITY AG, *ML5520va*. Available: <http://www.tridelly.com/de/displays/multi-view-3d-display-55%E2%80%B3-landscape/> (2016, Mar. 08).
- [17] ITU-T Rec. P.910 (04/2008), Subjective video quality assessment methods for multimedia applications. Retrieved from ITU-T.
- [18] Stereo Optical, Randot® Stereotest. Available: <http://www.stereo-optical.com/shop/stereotests/randot-stereotest/> (2016, Mar. 03).
- [19] Ishihara color charts. Available: http://www.biologiedidaktik.at/Humanbiologie/Dateien/Auge/ishihara_farbtafeln.pdf (2016, Mar. 03).
- [20] R. Kennedy, N. Lane, K. Berbaum, and M. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," Int. J. Aviation Psychology 3(3), pp. 203-220, 1993.