A study on order effect in a subjective experiment on stereoscopic video quality

Dawid Juszka and Zdzisław Papir
Department of Telecommunications
AGH University of Science and Technology
Kraków, Poland
Email: juszka@kt.agh.edu.pl

***Abstract*—Randomization is the best method to avoid any effects on the grading of tiredness or adaptation in a subjective experiment. Researchers' attention usually concentrates on a random order of stimuli in a clip displayed to a subject, but sometimes, due to the large number of stimuli to assess, it is necessary to divide the test material into a few clips and conduct a series of assessment sessions for each subject. Herein we present a study providing on evidence that there are significant differences in subjective assessment depending on the order in which clips are displayed, especially when the experiment concerns assessing the subjective experience of emerging technologies.**

I. Introduction

Each telecommunications operator is highly interested in offering the optimal quality of service. The main motivation is to reconcile user satisfaction with available resources. One of the methods of assessing the quality of multimedia services are subjective video quality experiments. Opinions collected from subjects can be applied to design quality metrics which can be an invaluable source of information on service performance from the end-user experience. That is why precision of collected scores is very important. Emerging types of services provide a previously unseen experience, so subjects taking part in an experiment have a very poor reference point, or even none at all. This is the main difference in the investigation process between novelties (i.e. stereoscopic video, augmented reality multimedia) and variations of previous, commonly available technologies (i.e. upgrade in resolution of 2D services). It is suspected that the lack of previous experience with a specific type of technology can bias the opinion scores.

An example of a modern multimedia service is the 3D Video on Demand platform. Indeed, what distinguishes subjective assessment experiments on stereoscopic video quality from 2D video quality is the novelty of 3D experience. Even now - 5 years after the first presentation of new 3D TV sets, it can be assumed that subjects taking part in the experiment are more familiar (have a wider reference knowledge) with 2D video than with stereoscopic video displayed in a non-cinema theatre environment. This fact is the main reason why experiment methodology should be chosen very carefully.

The experiment described below was conducted in 2011, only a year after the premiere of the *Avatar* movie (directed by James Cameron). Indeed, the audience in Poland (all subjects in the experiment were Polish) had already had their very first experience of the new generation of stereoscopic images, but did not experience it very often. Consequently, it was justified to be afraid that results might be biased by the halo effect or at least that the subjects' opinions could be unstable.

In this paper, a study is presented on the order effect in displaying clips in sessions (by clips we mean randomly concatenated stimuli). The main aim was to test the hypothesis that in a three-session experiment clips are scored by subjects significantly differently, depending on the order of clip presentation, against an alternative one (that order of clip display is not of significant importance).

The rest of the paper is structured as follows. Section 2 presents the related work in this area. It is followed by Section 3 which describes the test material and its preparation. Section 4 is a description of the psychophysical experiment. The results of the experiment are discussed in Section 5, which is followed by conclusions in Section 6.

II. Related work

Researchers who conduct subjective experiments encounter a lot of difficulties caused by human factors influencing the stimuli ratings. However, human visual perception is relatively objective when we consider its means (sight) and constant in terms of properties of the object being watched (light, surface, textures), but due to various factors residing in the observer's mind, it is significantly subjective. Obviously, observations need to be interpreted and this process is a complex phenomenon, unique to every individual. Consequently, visual perception is significantly influenced by internal and external subjective factors that are hard to grasp [1].

Such disciplines as psychology or neuroscience support the efforts to recognise and systematize the knowledge about biases originating from human factors. From an array of cognitive biases which influence evaluation processes in a way that can jeopardize the results of an experiment on up-to-date technology designs, two can be pointed out as the main suspects - order effect [2] and halo effect (one of the fundamental attribution errors) [3].

There are some clues, which lead to the conclusion that ITU-T Recommendation BT.500 [4] provides solutions to minimise such effects. Therefore, in the instructions for the assessment, it is stated that assessors should be carefully introduced to the method of assessment, the types of impairment or quality factors likely to occur, as well as the grading scale, the sequence and timing. In the beginning, there should be a training sequence presentation to demonstrate the range and type of impairments to be assessed. It should be used with illustrating pictures other than those used in the test, but of comparable sensitivity. Additionally, a random order should be used for the presentations; but the test condition order should be arranged so that any effects on the grading

978-1-4799-8958-4/15/$31.00 ©2015 IEEE

of tiredness or adaptation are balanced out from session to session. Some of the presentations can be repeated from session to session to check coherence. Moreover, in Appendix 3 to Annex 1 some recommendations raise the awareness of contextual effects which occur when the subjective rating of an image is influenced by the order and severity of impairments presented.

Another issue was found by Jumisko et. al. – their experiment has shown that evaluators with previous knowledge of the genre are more demanding regarding the acceptability of quality. So, familiarity with the content biases subjective scores – familiar content collected lower ratings than unfamiliar content [5].

Understanding and applying the knowledge of human perception set new requirements for quality evaluation methods. There are many human factors that are still under recognition and one of them is order effect. Not many efforts in this topic have been documented in the area of Quality of Experience for multimedia services, so best practise is based on researcher experience, intuition and a widely accepted methodology used in behavioural sciences.

III. Experiment design

A. Source Video Sequences

The source video sequences were produced by NTT (Nippon Telegraph and Telephone Corporation, Japan) and shared with the VQEG (Video Quality Experts Group) partners. From the available sequences the following were selected for further processing: “okugai”, “okunai”, and “digest”. Examples of frame sequences are presented in Fig.1.

Each stereoscopic source sequence was stored as a set of

Fig. 1. Frames from source sequences – first row: digest; second row: okuagi; third row: okunai

files, each file contained one frame for one eye (left or right view). Frames were stored in uncompressed TIFF format with 16 bit per channel colour resolution and 4k resolution, i.e. 3840×2016 pixels. Each file was 46.5 MB long. Video sequences were recorded with a progressive scan mode and 30 FPS (frames per second).

The sequences were cut into scenes. Only the most appropriate scenes were taken for the processing, including the 9 most diverse (in terms of content characteristics and content itself) 20-second long scenes. From this moment these 9 scenes will be called the SRCs (source video sequences).

Conversion from left and right 4k views to FullHD

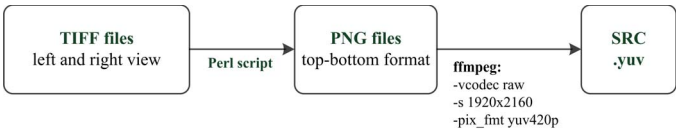

Fig. 2. Generation of SRCs.

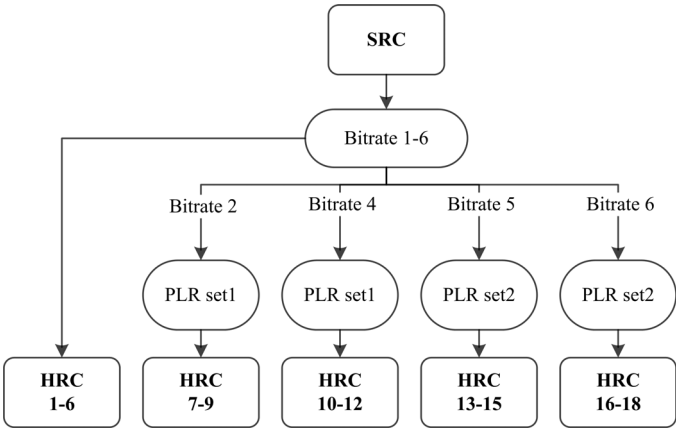

Fig. 3. Generation of HRCs.

”up-side down-side” lossless PNG format was realized using Perl script. The concatenated PNG files were 1920×2160 pixels: left and right view frames (1920×1080 pixels each) placed one over the other. In the last step of SRCs generation, single PNG files were transformed into video sequences in the YUV 420p format using the ffmpeg tool (see Fig. 2).
Since the conversion from 30 to 24 FPS (24 FPS is the limit of the HDMI 1.4 interface) produces unwanted artefacts, we have decided to preserve the original frame rate. It was possible thanks to the LCD BENQ XL2410T display which is capable of displaying stereoscopic video with FullHD resolution up to 60 FPS in the NVidia 3D Vision standard. In this case, a dual link DVI-D interface was used.

B. Hypothetical Reference Circuits (HRC)

The goal of the experiment was to analyse the influence of both compression and packet loss on stereoscopic FullHD video sequences. In order to prepare test material for the subjective experiment, the SRCs (source video sequences) were modified under 18 different HRCs (hypothetical reference circuits).

All SRCs were encoded into 6 different bitrates (HRC 1-6). Additionally, 3 levels of packet losses were added to 4 out of 6 considered bitrates, generating an additional $3x4 = 12$ conditions (HRC 7-18), see Fig 3 for details.
The SRCs were divided into 2 sets and each set was encoded into 6 different bitrates. The first set (SRC 1, 2, 4) was encoded into the following bitrates: 1500, 2000, 2500, 3000, 5000, and 50000 kbit/s. The second set (SRC 3, 5, 7, 8, 9, and 10) into 800, 1000, 1500, 2000, 5000, and 50000 kbit/s.
As presented in Fig. 3, two different sets of packet loss ratios were considered and applied to different bitrates. The following loss ratios were considered: 1) 0.2%, 0.75%, and 1% for set1, 2) 0.1%, 0.2%, and 0.75% for set2.

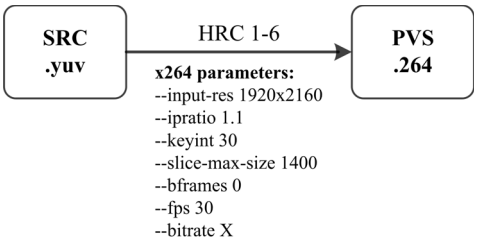

Fig. 4. Generation of PVSs including compression only.

C. Generation of Processed Video Sequences (PVSs)

The first step of the generation of the PVSs (processed video sequences) was the compression of the SRCs stored in a YUV format. The compression was a two-step process. Each SRC was encoded using x264 video codec into H.264 Annex B bitstream (see Fig. 4) with compression parameters as presented in Fig. 3. These files were used as an input for the packet loss generation chain details in the following section. In the second step, an AVI file container was added using ffmpeg software. The obtained video files represent the PVSs containing compression artefacts only (other PVSs contain also artefacts caused by packet losses).

D. Packet Loss Generation Chain

The set of previously prepared PVSs had to be streamed in order to add packet losses. For this purpose, we utilized an open source modular multimedia streaming software, called Sirannon [6]. In order to ensure fully controllable transmission (especially for the highest considered bitrate - 50000 kbit/s), the experiments were carried out between two separate PCs connected directly with an Ethernet cable. The server was established on Ubuntu 10.10 operating system with an appropriate software preinstalled and the client was equipped with Microsoft Windows 7 operating system. The RTP streaming of video files described in the RFC3984 document was chosen [7]. Input files for the server were XML files and the PVSs. The server structure was composed of an AVC-reader, AVC-packetizer, basic-scheduler and RTP-transmitter. All of the mentioned blocks were defined by previously generated XML files. The Wireshark software was run at the client side in order to capture the incoming video traffic. The buffer size was set to 50MB to make sure that no packets were dropped on the client network interface. The traffic was captured in the PCAP file format.

Packet losses were introduced to PCAP files using the PCAPLossGenerator tool [8]. The considered packet loss ratios have already been discussed in the previous sections. This entire process was realized using a Bash script. Video files in the H.264 Annex B bitstream format were extracted from the distorted PCAP files using H264AnnexBExtractor [8]. In the last step of the packet loss generation chain, an AVI container was added using the FFmpeg tool [9]. The process is illustrated in Fig. 5.

E. Concatenated Clips

Due to a large number of PVSs to be assessed by each subject, the repertory of all PVSs (156 items) had to be divided

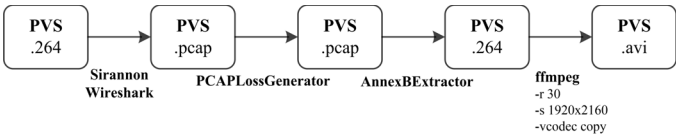

Fig. 5. Generation of PVSs including compression and packet losses.

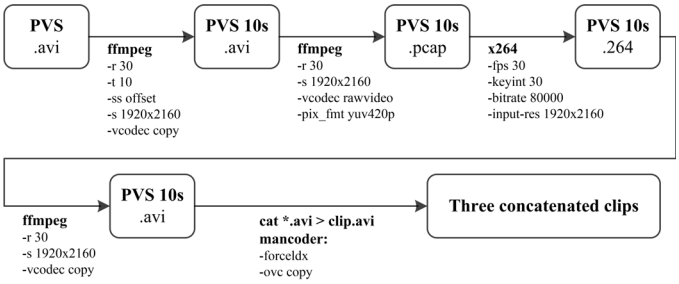

Fig. 6. Generation of PVSs including compression only.

into three parts (52 PVSs in each part) and each part was concatenated into one clip. Each clip contained different PVSs arranged in random order, but each SRC was represented in each clip. Each PVS was 20 seconds long (the same as the SRC) and had to be shortened to a 10-second sequence in order to fit the subjective experiment requirements. Within a single clip, all PVSs were interlaced with 15-second-long still image sequences informing about the voting for the recently viewed PVS and about the number of the subsequent PVS to be viewed. So, each session was about 21 minutes long (due to the novelty of the 3D experience, possible fatigue or discomfort that results from long clips should be accepted). Consequently, during the subjective experiment, in order to assess each PVS, each subject had to attend three sessions, during each session one of the three clips was presented. At the end, to minimise order effect, two trios of clips were prepared (first trio: clip 1, clip 2, clip 3; second trio: clip 4, clip 5, clip 6). The entire process of generating the concatenated clips is presented in Fig. 6.

IV. Experiment Setup

A. Test Environment Description

The test environment was arranged according to special requirements for conducting subjective experiments included in ITU-R Recommendation BT.1438 [10] (which corresponds with the well-known ITU-R Recommendation BT.500 [4]), ITU-T Recommendation P.910 [11] and IRU-R Recommendation BT.2021 [12]. A device deployed to present the test material was a 24" BENQ 3D display with shutter glasses. The viewing distance for the subjects was set by the subjects themselves, the range was between 0.75 and 1 meter. Environmental illumination was set to approximately 200lx, which is an appropriate value [4]. The test room is presented in Fig. 7.

B. Subjects

Subjects were selected by a recruitment agency which pre-checked their stereo vision acuity, using the Randot Stereo Test. A positive result was a necessary condition for taking part in the experiment. All of the 30 subjects were naive observers,

Fig. 7. Test environment

meaning that their profession was not connected with television picture quality or assessment of services. Each subject had either normal vision acuity or was supported with proper corrective glasses or contact lenses. In the laboratory room each subject was examined with the Ishihara colour vision test and visual acuity test with Snellen chart. All subjects were paid to take part in the experiment. Statistical information about the participants is as follows:

- sex: 16 men, 14 women;
- age: average 31.3 years; median 28.5 years; minimum age limit 18 years; maximum age limit 51 years;
- education: secondary education – 13 participants; higher education (college or university) – 17 participants.

C. Methodology Description

The main reason for conducting this experiment was to gather results of human perceived quality, so subjects taking part in it were asked to rate the experienced video quality of the presented 3D video sequences. A 1-5 categorical quality scale was used as a rating tool, where 1 represents poor quality perception and 5 represents excellent quality perception,. The subjective test methodology selected as a scenario for this test was ITU's ACR (Absolute Category Rating, described in ITU-T P.910 [11]).

Each subject took part in a training session to get familiar with the test procedure and then in three sessions separated with a break of a few minutes. During each session one clip was presented. Additionally, being aware of any effects on the grading of adaptation or tiredness, the order of clips presented to the subjects (concerning their gender) was also randomized, i.e. clip 1 was displayed to the first female subject during her first session, but to the second female subject the same clip was presented during her third session. The order of clip presentation during sessions for each subject is presented in Tab. I

TABLE I. ORDER OF CLIP PRESENTATION DURING SESSIONS FOR EACH SUBJECT.

Subject ID	1st session	2nd session	3rd session
1	1	2	3
2	1	3	2
3	2	1	3
4	2	1	3
5	1	2	3
6	2	1	3
7	2	3	1
8	1	2	3
9	2	3	1
10	3	2	1
11	1	3	2
12	2	1	3
13	3	1	2
14	1	3	2
15	2	3	1
16	3	1	2
17	5	4	6
18	6	4	5
19	6	5	4
20	4	6	5
21	5	4	6
22	6	5	4
23	4	5	6
24	5	6	4
25	6	4	5
26	4	6	5
27	4	5	6
28	5	6	4
29	4	5	6
30	6	5	4

V. DATA ANALYSIS AND RESULTS

Before analysing the scores provided by the subjects, we first screened the outliers according to the subjective scores given by the subjects. This process was based on the guidelines provided in section 2.3.1 of annex 2 of ITU-R Recommendation BT.500 [4]. Three outliers were identified – subjects with ID numbers 1, 11 and 25. Two of them can be suspected of working under the halo effect – no matter which PVS was presented one of them always scored 1 or 2, and the second one – 4 or 5.

We used the Mann-Whitney U test to test the research hypothesis that in a three-session experiment, PVSs (concatenated in clips) are scored by subjects significantly higher (or lower), depending on the order of clip presentation against an alternative one (that the order of clip display is not of significant importance). This test is a nonparametric alternative to the t-test for independent samples, the results of which are shown in Tab. II. From the collected data, pairs were selected for comparison. For example, scores collected for clip 1 presented to subjects during their first session of the experiment were compared with scores collected when clip 1 was presented in the second session (that is why in the second row in Tab. II there is an X in the cell representing the sum for the third session).

Definitions for each statistic from this table are as follows:

$$U = R_i - \frac{n_i(n_i - 1)}{2} \tag{1}$$

For large samples, statistic U is approximately normally distributed. In that case, the standardized value:

$$Z = \frac{U - \mu}{\sigma} \tag{2}$$

TABLE II. COMPARISON OF MOS SCORES OF THE SETS PRESENTED IN DIFFERENT ORDER.

	1st session rank sum	2nd session rank sum	3rd session rank sum	U	Z	p	Z*	p
clip 1	53866	81594	X	32130	-0.19	0.85	-0.19	0.85
clip 1	45764.5	X	40971.5	19235.5	1.95	0.05	**2.02**	**0.044**
clip 1	X	85440.5	50019.5	28283.5	2.48	0.01	**2.56**	**0.01**
clip 2	98885.5	36574.5	X	24328.5	2.59	0.01	**2.67**	**0.008**
clip 2	97587.5	X	66290.5	31157.5	-3.52	0.0004	**-3.63**	**0.0003**
clip 2	X	24032	42398	11786	-4.47	0.000008	**-4.57**	**0.000005**
clip 3	36536.5	49368.5	X	15698.5	3.71	0.0002	**3.83**	**0.0001**
clip 3	42471.5	X	66806.5	17978.5	4.51	0.000006	**4.64**	**0.000004**
clip 3	X	76542.5	86763.5	37935.5	1.26	0.2	1.29	0.196
clip 4	56339	29981	X	17891	1.91	0.06	1.96	0.05
clip 4	65151.5	X	70308.5	31221.5	-1.51	0.13	-1.55	0.12
clip 4	X	28447.5	57872.5	16357.5	-3.21	0.0013	**-3.29**	**0.00099**
clip 5	52838	82102	X	31310	-0.59	0.56	-0.61	0.55
clip 5	38905	X	27161	14915	1.24	0.21	1.27	0.2
clip 5	X	75952	33794	21548	2.02	0.04	**2.084**	**0.037**
clip 6	43659	42246	X	20510	0.75	0.45	0.77	0.44
clip 6	45853.5	X	62491.5	24532.5	-1.49	0.14	-1.54	0.12
clip 6	X	45226	64052	23490	-2.377	0.0175	**-2.459**	**0.0139**

where μ and σ are the mean and the standard deviation of U, is approximately a standard normal deviate whose significance can be checked in tables of normal distribution. μ and σ are given by:

$$\mu = \frac{n_1 n_2}{2} \tag{3}$$

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{4}$$

For a two-tailed test with α = 0.05, the null hypothesis must not be rejected if

$$-1.96 < z* < 1.96 \tag{5}$$

Calculations presented in Tab. II show that the inequation is true only for 8 out of 18 comparisons. Consequently the test hypothesis in these cases should not be rejected, but since p-values for each of these eight cases are higher than 0.05, these results are not statistically significant. For the remaining 10 out of 18 comparisons (bolded in Tab), that inequation is false, so it suggests that real differences in the assessment do exist. What is more, p-value for each of these cases is lower than 0.05, so the results are statistically significant.

VI. CONCLUSIONS AND FUTURE RESEARCH

A subjective experiment on stereoscopic video quality assessment was presented. At the time when the experiment was conducted, stereoscopic video was a novel kind of multimedia, so consumers did not have wide experience of it in their day-to-day life. Two types of impairments were introduced - compression and packet loss. Thirty subjects took part, each assessed test material (divided into three clips) in three sessions separated by short breaks. The study shows that subjective assessment of the overall quality of video sequences depends on the order of clip presentation - collected opinion scores do differ significantly. This observation suggests that researchers should be aware of order bias when designing subjective experiments. It is interesting that for each clip (excluding clip 3) a comparison of scores for the second and third session shows a significant difference. Additionally, for each clip a comparison of subjective assessment results for the first and third session differ, but not always significantly. This observation leads to the conclusion that subjects learn a scale during the task or become more experienced in stereoscopic video perception. Problems with scale learning in subjective experiments on emerging technologies can probably be solved by employing the Pair Comparison method (instead of ACR). However, this suggestion needs scientific confirmation. Further investigation is needed to measure the effect size of clip order.

ACKNOWLEDGMENTS

The authors would like to thank Jarosław Bułat, Michał Grega, Lucjan Janowski, Mikołaj Leszczuk, Błażej Szczerba and Piotr Romaniak for their help in preparing the above described experiment. This work has been supported by the Dean's Grant (agreement number 15.11.230.149) - Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology.

REFERENCES

[1] M. Mirkovic, P. Vrgovic, D. Culibrk, D. Stefanovic, and A. Anderla, "Evaluating the role of content in subjective video quality assessment," *The Scientific World Journal*, vol. 2014, Jan. 2014.

[2] P. C. Cozby, *Methods in Behavioral Research, 10th Edition*, 2008.

[3] R. Corsini, *The Dictionary of Psychology*. Routledge, 2001.

[4] ITU-R, *Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures*, International Telecommunication Union, Geneva, Switzerland, 2012.

[5] S. H. Jumisko, V. P. Ilvonen, and K. A. Väänänen-vainio mattila, "Effect of TV Content in Subjective Assessment of Video Quality on Mobile Devices," in *Proc. of SPIE-IS&T Electronic Imaging*, R. Creutzburg and J. H. Takala, Eds., vol. 5684, Mar. 2005, pp. 243–254.

[6] A. Rombaut, *SIRANNON 0.6.10: Modular Multimedia Streaming*, University of Ghent, IBCN., May 2011, http://sirannon.atlantis.ugent.be/.

[7] S. Wenger, M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, *RTP Payload Format for H.264 Video*, The Internet Society, February 2005, http://www.ietf.org/rfc/rfc3984.txt/.

[8] B. Szczerba and D. Ziobro, *Generating of packet loss in the video sequences encoded with H.264 video codec*, AGH University of Science and Technology, Atlanta, USA, November 2010, ftp://vqeg.its.bldrdoc.gov/Documents/.

[9] *FFmpeg*, http://ffmpeg.org/.

[10] ITU-R, *ITU-R Recommendation BT.1438, Subjective assessment of steroscopic television pictures*, International Telecommunication Union, Geneva, Switzerland, 2000. [Online]. Available: http://www.itu.int/rec/R-REC-BT.1438-0-200003-I

[11] ITU-T, *ITU-T Recommendation P.910, Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union, Geneva, Switzerland, 1999. [Online]. Available: http://www.itu.int/rec/T-REC-P.910-200804-I

[12] ITU-R, *Recommendation IRU-R BT.2021: Subjective methods for the assessment of stereoscopic 3DTV systems BT Series Broadcasting service*, International Telecommunication Union, Geneva, Switzerland, 2012.