https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-527
© 2018, Society for Imaging Science and Technology

Exploring the effects of subjective methodology on assessing visual discomfort in immersive multimedia

***Jing Li*[†]*, Junle Wang*[♯]*, Marcus Barkowsky*[†]*, Patrick Le Callet*[†]**
[†] *IPI/LS2N Lab, University of Nantes, France*
[♯] *Ars Nova Systems*

Abstract

Visual discomfort is an important factor that influences viewing experience in immersive multimedia, for example, 3DTV and VR. With the added value of depth, the novel perceptual experience, visual discomfort is not an easy task for observers to evaluate. In this study, we investigate how the subjective methodology affects the test results in 3DTV condition. Two subjective visual discomfort experiments were conducted. One used the Pair Comparison (PC) method and the other used the Absolute-Category Rating (ACR) method. The results demonstrated that PC method had more powerful discriminability. For a difficult perceptual-related tasks, such as visual discomfort in our study, PC was more easy to understand and conduct for the observers which led to reliable results. It also showed some very important but usually ignored conclusions on the subjective experiment, i.e., for measuring the perceived visual discomfort, the observer's judgment behavior might be affected by the test methodology.

Introduction

As one of the most important dimensions in Quality of Experience (QoE) of immersive multimedia, visual discomfort is often complained by the viewers. Thus, recent researches are concentrated on the possible causes of visual discomfort, e.g., the vergence-accommodation (VA) conflict[1][2], the excessive binocular disparity[3], the relative disparity between foreground and background[4][5], the motion[6][7][8], crosstalk[9][10]], etc. Another research direction is focusing on the development of objective prediction methods to automatically monitor, adjust or optimize the related systems and thus, to minimize the possibility of visual discomfort[11][12][13][14][8]. Nevertheless, it should be noted that the fundamental of these studies is subjective experiment.

The complexity of stereoscopic content perception as opposed to real-world perception explains the difficulties that naive observers experience when asked to provide an opinion on the visual experience. On one hand they have limited experience with the new technology, notably as opposed to 2D television and, eventually, immersive multimedia content. On the other hand, they may need to counterbalance positive and negative effects such as added depth value and visual discomfort.

In ITU-R BT.2021 [15], four assessment methods are recommended for measuring visual discomfort, which are a subset of the methods from Recommendation ITU-R BT.500 [16]. These four methods are:

- the single-stimulus(SS) method;
- the double stimulus continuous quality scale (DSCQS) method;
- the stimulus-comparison (SC) method;
- the single stimulus continuous quality evaluation (SSCQE) method.

Compared with the conventional 2D quality assessment scale labels, the labels for visual discomfort are slightly different, for example, the discrete five-grade scales or the continuous comfort scales are labeled with "Very comfortable", "Comfortable", "Mildly uncomfortable", "Uncomfortable", and "Extremely uncomfortable".

These methods have already been widely used in the community of Stereoscopic 3DTV. For example, in [17], the SSCQE method was used as it can measure the influence of stimulus duration on visual discomfort or visual fatigue. In [7], a continuous scale from 0 to 100 was used, where "0" represents "Extremely Uncomfortable" and "100" represents "Very Comfortable". In [6], five scale based Absolute-Category Rating (ACR) methodology was used, where the score from 1 to 5 represents "I'm very tired" to "I am not tired". While in [14] a 5-point ACR test was used as well, the attributes were selected from "very comfortable" to "extremely uncomfortable".

Pair Comparison (PC) is considered as a more powerful test methodology recently. Compared to the scale-based subjective methodology, PC is easy to interpret by testers and to understand by observers [18]. In addition, PC outperforms the SS methodology in terms of the discriminability in image quality [19][20].

So far, there are few studies on the comparison of the visual discomfort results obtained by different test methodologies with or without the same test conditions. Due to the multi-dimensionality of the QoE, and the difficulties for the viewers to make judgment on unfamiliar and multi-dimension scales, it would be interesting to know the influence of the test methodology and test condition on results.

In this study, two visual discomfort experimental results on the same video database are compared. One experiment was conducted by the 5-point ACR method in [14], and the other was conducted in IPI lab using the ORD (*Optimized Rectangular Design*) PC method[21][22] recommended by ITU-T P.915[23] and IEEE P3333.1.1[24]. More details are presented in the following sections.

Test Stimuli

In this study, the IVY Lab stereoscopic video database [25] is chosen as it contains different types of motion. This database includes 40 video sequences, and 36 of the video sequences were shot by the IVY lab using the Fujifilm FinePix 3D W3 camera with dual lenses, the remaining 4 are video sequences from the MPEG 3D video test. In order to avoid the effect of excessive binocular disparity on visual discomfort, the maximum disparity of the sequences is within the comfortable viewing zone (1 degree). The motion types include vertical planar motion, horizontal planar motion, in-depth motion and their combinations. The horizontal motion velocity ranges from 1.83 to 25.5 degree/s. The vertical motion velocity is ranged from 0.05 to 3.37 degree/s, and the depth motion velocity is ranged from 0.05 to 3.37 degree/s. The motion and disparity are estimated by an 8 × 8-pixel block matching method [25]and the depth estimation reference software (DERS from MPEG 3D video standardization) [26], respectively. The resolution of the video sequences is 1280 × 720, and the frame rate is 24 fps. The duration of each sequence is 10 seconds.

In this study, we only chose the 36 stimuli which were shot by the IVY Lab. The reasons were that firstly, they were shot in the same shooting conditions while the remaining 4 MPEG 3D video test sequences were not. Furthermore, considering the test duration for PC test, using 36 stimuli is feasible for applying the ORD method and it already reaches the maximum limit for test duration, approximately 1 hour (180 pairs = 180×(10+5)s = 45 minutes without break). Using 40 stimuli would make the test even longer (for 8× 5 condition, the total number of pairs = 220 pairs = 220×(10+5)s = 55 minutes without break), which is not recommended.

A preview of the video sequences used in the subjective test is shown in Figure 1. Please note that the indices of the video sequences are consistent with the original IVY Lab database, the video sequences 1, 21, 22 and 40 are the MPEG 3D video test sequences which were not chosen in this study.

Experiment

Experiment 1: ACR test conducted at the IVY lab

It should be noted that Experiment 1 is not our work but the original work of IVY lab [25]. It is briefly introduced here for easier comparison between the experiment in IVY lab and the experiment in our lab.

Apparatus

A linearly polarized stereoscopic monitor manufactured by Redrover (true3Di) was used in the test. It consisted of a half mirror and two 40” LCD displays with the refresh rate of 60 Hz. The width and height of the display screen were 886 mm and 498 mm, respectively. The resolution of the screen is 1920 × 1080. The viewing distance was approximately three times of the height of the screen, i.e., 150 cm. In the test, when displaying the video sequence, the original video (1280 × 720) was re-scaled to fit the full screen. The test environment was in line with the recommendations of ITU-R BT.500 [16].

Viewers

17 subjects, aged from 20 to 37 years old, participated in the test. All subjects were recruited under approval of the KAIST Institutional Review Board. All subjects had normal or corrected vision and a minimum stereopsis of 60 arcsec in stereo fly test.

Test methodology

In the subjective experiment, the ACR method was used to get the **visual comfort** scores, the 5-point scale values represent:

- 5: very comfortable (visual discomfort is imperceptible)
- 4: comfortable (visual discomfort is perceptible but not annoying)
- 3: mildly uncomfortable
- 2: uncomfortable
- 1: extremely uncomfortable

Between each two video sequences, there is a resting time of about 15s with mid-gray image. During the resting time, observers were asked to provide an overall level of **visual comfort** for the tested video sequence.

Experiment 2: PC test conducted at the IPI lab

To compare the experimental results between the ACR and PC methods, a PC test was conducted in our IPI lab with the experimental setup as close to Experiment 1 as possible. To reduce the number of comparisons, our proposed ORD method was used[21][22] . Details are shown in the following sections.

Apparatus

Two ViewSonic V3D231 (model number: VS14136) polarized display were used in the test. They were positioned side by side. The size of the screen is 23”, with resolution of Full HD (1920×1080). The refresh rate is 60 Hz. To conform to the conditions used in the IVY lab, in our test, when displaying the video sequence, the original video was re-scaled to fit the full screen. Viewing distance was about 3 times of the screen height, i.e., 87 cm. The display was adjusted for a peak luminance of 210 cd/m^2, approximately 80 cd/m^2 through polarized glasses. The background illumination was about 30 cd/m^2, approximately 12 cd/m^2 through the polarized glasses. All other environmental conditions were in line with ITU-R BT.500 [16]. This setup was consistent with the experiment conducted in IVY lab besides the size of the screen.

Viewers

40 naive viewers participated in this test. 22 are females and 18 are males. Their ages were ranged from 19 to 65, with an average age of 30.2. All of them had either normal or corrected-to-normal visual acuity. The visual acuity test was conducted with a Snellen Chart for both far and near vision. The Randot Stereo Test was applied for stereo vision acuity check, and Ishihara plates were used for color vision test. All of the viewers passed the pre-experiment vision check.

Test methodology

As there were in total 36 video sequences, and the MOS from Experiment 1 was already available, the ORD (*Optimized Rectangular Design*) method [21][22] was used, i.e., the stimuli with closer visual discomfort would be arranged in the same column or row of a matrix, and only the stimuli in the same column or row would be compared. Thus, the square matrix in ORD is arranged based on the rank ordering of the MOS as shown in Table 1 (the number in the matrix represents the index of the stimulus). The video sequences with the closest visual discomfort MOS were put in the same column or row thus they will be directly compared. This direct comparison on closest pairs allows for a precise preference evaluation between the MOS scores and PC binary data. Conforming to the ORD method, there were in total 180 pairs for

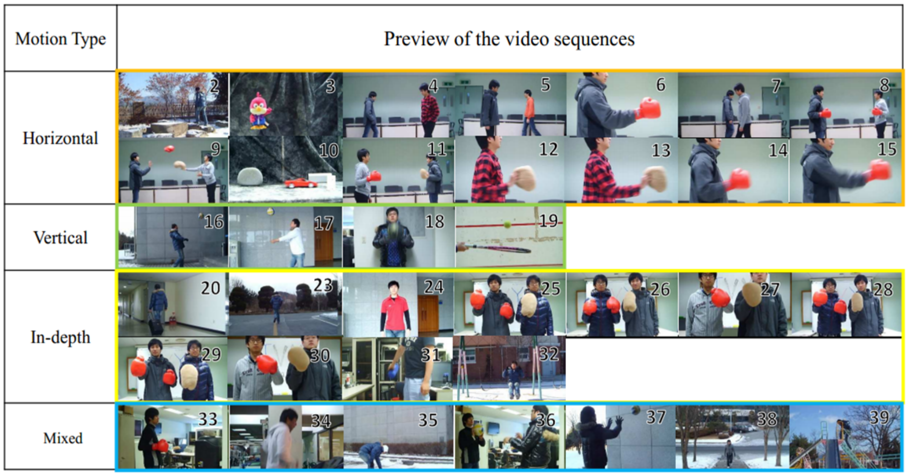


Figure 1. *Preview of the test video sequences. They are captured from the 100th frame.*

Table 2: The square matrix used in the PC experiment in IPI lab. The number in the matrix represents the index of the stimulus. Only the stimuli in the same column or row were compared.

19	15	28	31	10	13
4	35	9	17	36	14
34	24	23	38	37	29
11	8	20	39	3	25
12	2	16	6	18	30
33	7	27	5	26	32

each observer.

Procedure

The test included a training session and a test session. Five pairs were included in the training session. After watching a pair of video sequences, the observers were asked to select the one which is more comfortable. A touch screen was used for the viewers to make the selection. If the observer was not very sure about the selection, he/she could replay the video sequences as many times as he wanted.

There were in total 180 pairs for each viewer. The video pairs were randomly presented to all viewers. Meanwhile, the presentation order for each viewer and all observers were as balanced as possible, which meant the video sequence should be presented with the same frequency on the left screen and on the right screen. For the sequence pair $\{AB\}$, the presentation order of $\{A\text{-}B\}$ should appear as often as the condition $\{B\text{-}A\}$ for all observers. In this way, the presentation bias effect was avoided as much as possible.

Each test session was split into two sub-sessions. After half of each sub-session, the viewers were asked to have a 10 minutes break to avoid visual fatigue. When finishing the first sub-session, the screen showed a message saying “End of the first session” to the viewers. The viewers could take a break and then pressed the “continue” button to move to the second sub-session. The whole test lasted approximately 1 hour.

Results: Comparison between ACR and PC

The results obtained in Experiment 1 and Experiment 2 are compared in this section. The differences between the two results are analyzed based on two main aspects. One is focusing on the scale values after being converted from the raw PC data. The other is focusing on the raw PC data.

It should be noted that to make a fair comparison between ACR and Full PC, they should be with the same observers, i.e., ACR test with 17 observers versus Full PC test with 17 observers. Considering that our PC method is an efficient design (with reduced number of comparisons), to make the comparison fair, our ORD PC test should have same comparison number with the Full PC of 17 observers, which leads to (36x35/2)x17 = 10710 comparisons. This means 10710/180 = 59.5 observers in our ORD PC test. However, we did not have so many observers in the test, thus, we only take 40 observer’s results to compare (which is in fact unfair for ORD PC results).

The MOS and 95% confidence intervals for all sequences from Experiment 1 are available from the website [25]. The Bradley-Terry (BT) model[27][28] is used to generate visual comfort scores of Experiment 2. The higher the BT score, the higher the degree of visual comfort. It should be noted that in IVY lab, the MOS also represents the degree of visual comfort[14].

Comparison between the scales values: MOS and BT scores

The scatter plot of the MOS and BT scores is shown in Figure 2. The CC, SROCC, RMSE between the MOS and BT scores are calculated, which are 0.53, -0.50 and 0.33, respectively.

For convenience, the scatter plot based on the types of motion is provided in Figure 3. As shown in Figure 2, the correlation between the MOS and BT scores for the mixed motion are higher than the other conditions. In particular, for the condition of in-depth motion sequences, the BT scores are significantly different but overlapped on MOS.

Furthermore, it is shown that the confidence intervals of the

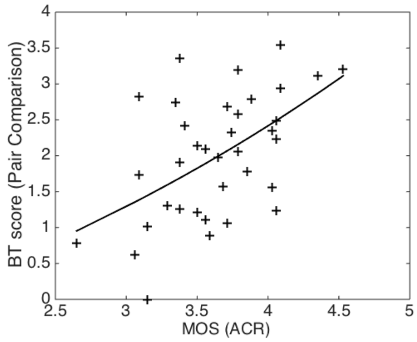
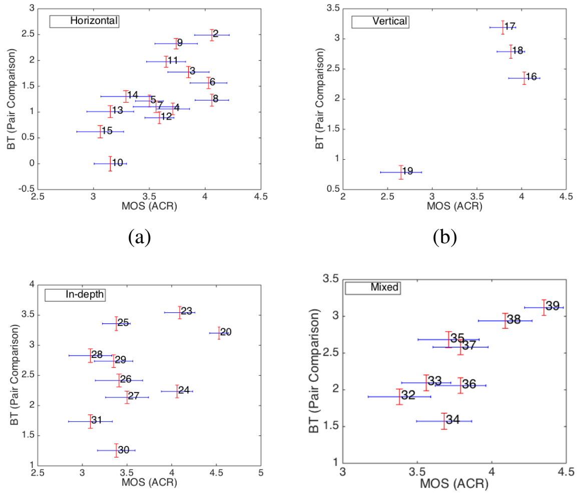


Figure 2. *The scatter plot of the MOS results and BT scores. The black line is the fitting curve from MOS to BT scores.*



(a)

(b)

(c)

(d)

Figure 3. *The scatter plot between BT score and MOS based on the type of motion, the error bar shows the 95% confidence intervals. The number labeled close to the marker represents the video index.*

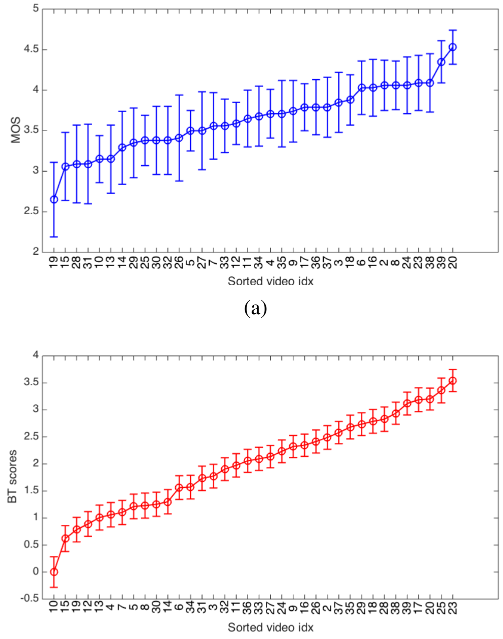
MOS are larger than the BT scores. For better visualization, the sorted MOS and BT scores are shown in Figure 4. For MOS, they are ranged from 2.5 to 4.5. According to the confidence intervals, a large amount of the scores are not significantly different. For example, the confidence intervals for the video sequence 15, 28, 31, 10, 13, 14, 19 are overlapping. On the contrary, for the BT scores, the number of the overlapping confidence intervals is smaller. To better evaluate the viewers' agreement on the scores, some statistical analysis are applied on the raw data, which will be introduced in the following section.

Comparison of the raw data

In this section, the obtained raw data from ACR and PC experiment are compared using different analysis methods.

Discriminability test

To compare the discriminability of the MOS and the PC data, the Barnard's-exact test is applied on the PC data. The objective is to compare the discriminability of the ACR method and PC method.



(a)

(b)

Figure 4. *The comparison between the sorted MOS and BT scores.*

There are in total 35 adjacent pairs in MOS, for example, sequence{19,15}, sequence{15, 28}, ..., sequence{39, 20} (as shown in Figure 4(a)). According to the confidence intervals of these adjacent pairs, the MOS of the stimuli in each adjacent pair are not significantly different. To evaluate their significance in the PC test, the Barnard's test on the preference of these adjacent pairs are calculated. The pairs whose p-value $\leqslant 0.05$ (which indicates a significant difference between the votings on the video sequence A and B at the significance level of 0.05) are shown in Table 2. According to Table 2, 20 out of 35 pairs are significantly different.

To provide more detailed information about the discriminability of the two test methodologies, all 180 pairs were tested by Barnard's test. Meanwhile, the significance test on the corresponding 180 pairs of the ACR results were conducted by using the student's-t-test. For better understanding, in this test (see Table 3), "PC1_ACR0" is used to represent the number of pairs that PC succeeds in detecting their significant difference but the ACR test fails. Thus, "1" represents the method that succeed in detecting the significant difference, "0" represents failure. The same meaning applies to the notion "PC0_ACR0", "PC0_ACR1", and "PC1_ACR1". The test results are shown in Table 3. The results indicated that there are in total 27 pairs can be discriminated by the ACR method and 78 pairs can be discriminated by PC test. The number of pairs that discriminated by the PC method is approximately 3 times of the ACR method. Thus, it could be concluded that PC comparison method has higher discriminability than the ACR method on the visual discomfort induced by different video sequences.

This study verifies the conclusions from [19] that the PC method has higher discriminability on closer stimuli. In addition, the results showed that the test methodology may affect viewer's behavior during the test. For example, in our paired comparison

Table 3: Barnard's test results: The adjacent pairs of the MOS which show significantly difference in PC experiment, p-value $\leqslant 0.05$

Sequence A	Sequence B	Vote on A	Vote on B	Barnard's p-value
15	28	34	6	0.00
28	31	7	33	0.00
31	10	9	31	0.00
10	13	31	9	0.00
14	29	33	7	0.00
29	25	33	7	0.00
25	30	5	35	0.00
26	5	11	29	0.02
5	27	29	11	0.02
27	7	12	28	0.03
33	12	8	32	0.00
4	35	32	8	0.00
9	17	32	8	0.00
17	36	11	29	0.02
37	3	11	29	0.02
18	6	8	32	0.00
6	16	28	12	0.03
2	8	9	31	0.00
8	24	29	11	0.02
24	23	33	7	0.00

Table 4: Comparison between the discriminability of the ACR and PC test on visual discomfort of the video pairs.

PC0_ACR0	PC0_ACR1	PC1_ACR0	PC1_ACR1
75	12	78	15

test, the viewers might pay more attention on the effect of window violation than in the ACR test. However, the differences between the two test results are not only from the test methodologies, but also possibly from some other factors.

Observer agreement test

To evaluate the agreement between the observer's individual test result and the global results (i.e., MOS for ACR test, and all observers' combined PC raw data for PC test), an observer agreement test is conducted. For ACR test, the MOS is converted to a binary 36×36 PC matrix M_{ACR}. The value in $M_{ACR}(m,n)$ is used to represent if the MOS of sequence m is higher than that of sequence n. The observer's individual scale rating is converted to a binary matrix in the same way, which is denoted by $M_{ACRi}(m,n)$, i represents the observer ID. For PC test, the global observers' PC matrix is converted to a binary matrix, denoted by M_{PC}. The observers' individual PC matrix is obtained directly in the experiment, denoted by M_{PCi}.

For ACR experiment, the agreement test is to calculate the ratio that the value in each position of M_{ACR} equals to the corresponding value in M_{ACRi} for each observer. The same procedure is applied on the PC experiment. Figure 5 shows the histogram of the agreement ratio in two experiments. The mean value of agreement ratio in ACR test is 0.43, in PC test is 0.70.

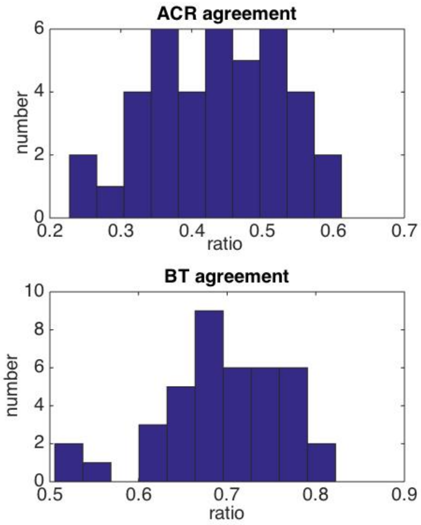


Figure 5. The histograms of the agreement ratio for ACR experiment and PC experiment. The X-axis represents the agreement ratio of individual observer and global results. The Y-axis represents the number of observers.

The agreement test results indicate again that in visual discomfort assessment, ACR test methodology is more difficult for observers to understand and conduct, so the obtained results are less consistent compared to PC methodology.

Observer behavior analysis

To analyze which factors lead to the big difference between the two test results, we checked the significantly different pairs between ACR and PC, as shown in Table 2, and the characteristics of these videos sequences. We noticed that most viewers' selections were concentrated on the video sequence that have window violation. Window violation is a phenomenon in 3D images or videos that when an object with strong crossed disparity (in front of the screen) interferes with the boundaries of the screens (bottom, top, left and right). In such cases, the object is perceived as being cut off by the borders. This unnatural shooting distortion would induce visual discomfort [29].

In the IVY stereoscopic video database, Sequence 6, 10, 12, 13, 14, 15, 19, 24, 27, 30, 31, 33, and 36 have window violation. Based on the results in Table 2, it might be inferred that when using the PC method, besides the large relative disparity and the motions, the window violation became a key factor for viewers to make the judgment, especially for the conditions that one stimulus had window violation while the other did not, such as the Stimuli pair {15, 28}, {28, 31}, {14, 29}, {25, 30}, {33, 12}, {17, 36}. However, in the results of ACR method, the effect of window violation might not be observed as in PC test because according to the confidence intervals of the MOS, the visual discomfort induced by these pairs are not significantly different. The ignorance of the window violation in ACR test showed the different mechanisms between PC and ACR on measuring visual discomfort.

Conclusion

To what extent the PC methodology is different from the ACR methodology in the context of visual discomfort in immersive multimedia is the question to be resolved in this study. In this paper, the visual discomfort results obtained by the ACR and PC test methodologies are compared. The results verified the conclusion that the PC method has higher discriminability than the ACR method. In addition, in a visual discomfort experiment, PC method is more easy to understand and conduct for the observers. People may feel confusing on the visual comfort scale in ACR test, which lead to unreliable and inconsistent results. It has also demonstrated that the viewer's behavior during the test might be influenced by the test methodology. The conclusions of this study are very important for the studies which utilize the subjective experimental results as the ground truth. The researchers should notice that the obtained results might not be the "ground truth" results and they might have been affected by the test methodology.

Acknowledgments

The participants of the subjective experiment are gratefully acknowledged. Many thanks to Prof. Jung for providing their original ACR subjective experiment data.

References

[1] D. Hoffman, A. Girshick, K. Akeley, and M. Banks, "Vergence–Accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision* **8**, pp. 1–30, Mar. 2008.

[2] J. Kim, T. Shibata, D. Hoffman, and M. Banks, "Assessing vergence-accommodation conflict as a source of discomfort in stereo displays," *Journal of Vision* **11**(11), pp. 324–324, 2011.

[3] S. Ide, H. Yamanoue, M. Okui, F. Okano, M. Bitou, and N. Terashima, "Parallax distribution for ease of viewing in stereoscopic hdtv," in *Electronic Imaging 2002*, pp. 38–45, International Society for Optics and Photonics, 2002.

[4] J. Li, M. Barkowsky, and P. Le Callet, "The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos," *International Workshop on Quality of Multimedia Experience* , pp. 155–160, Sep. 2011.

[5] J. Li, M. Barkowsky, J. Wang, and P. Le Callet, "Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses," in *17th International Conference on Digital Signal Processing (DSP)*, pp. 1–8, IEEE, 2011.

[6] S. Yano, M. Emoto, and T. Mitsuhashi, "Two factors in visual fatigue caused by stereoscopic HDTV images," *Displays* **25**(4), pp. 141–150, 2004.

[7] F. Speranza, W. Tam, R. Renaud, and N. Hur, "Effect of disparity and motion on visual comfort of stereoscopic images," *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems* **6055**, pp. 94–103, Jan. 2006.

[8] J. Li, M. Barkowsky, and P. Le Callet, "Visual discomfort of stereoscopic 3d videos: Influence of 3d motion," *Displays* **35**(1), pp. 49–57, 2014.

[9] P. Seuntiëns, L. Meesters, and W. Ijsselsteijn, "Perceptual attributes of crosstalk in 3D images," *Displays* **26**(4-5), pp. 177–183, 2005.

[10] S. Pastoor, "Human factors of 3d imaging: results of recent research at heinrich-hertz-institut berlin," in *Proc. IDW*, **95**(3), pp. 69–72, 1995.

[11] Y. Nojiri, H. Yamanoue, S. Ide, S. Yano, and F. Okana, "Parallax distribution and visual comfort on stereoscopic hdtv," in *Proc. IBC*, pp. 373–380, 2006.

[12] D. Kim and K. Sohn, "Visual fatigue prediction for stereoscopic image," *Circuits and Systems for Video Technology, IEEE Transactions on* **21**(2), pp. 231–236, 2011.

[13] S.-i. Lee, Y. J. Jung, H. Sohn, and Y. M. Ro, "Subjective assessment of visual discomfort induced by binocular disparity and stimulus width in stereoscopic image," in *IS&T/SPIE Electronic Imaging*, pp. 86481T–86481T, International Society for Optics and Photonics, 2013.

[14] Y. Jung, S. Lee, H. Sohn, H. W. Park, and Y. Ro, "Visual comfort assessment metric based on salient object motion information in stereoscopic video," *Journal of Electronic Imaging* **21**(1), pp. 011008–1, 2012.

[15] ITU-R BT.2021, "Subjective methods for the assessment of stereoscopic 3DTV systems," *International Telecommunication Union, Geneva, Switzerland* , Aug. 2012.

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

[17] S. Yano, S. Ide, T. Mitsuhashi, and H. Thwaites, "A study of visual fatigue and visual comfort for 3D HDTV/HDTV images," *Displays* **23**(4), pp. 191–201, 2002.

[18] U. Engelke, Y. Pitrey, and P. Le Callet, "Towards a framework of inter-observer analysis in multimedia quality assessmnet," *International Workshop on Quality of Multimedia Experience* , pp. 183–188, Sep. 2011.

[19] J.-S. Lee, L. Goldmann, and T. Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia Tools and Applications* , pp. 1–18, Feb. 2012.

[20] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-d synthesized view assessment," *Selected Topics in Signal Processing, IEEE Journal of* **5**(7), pp. 1332–1343, 2011.

[21] J. Li, M. Barkowsky, and P. Le Callet, "Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs," *IS&T/SPIE Electronic Imaging* , Feb. 2013.

[22] J. Li, M. Barkowsky, and P. Le Callet, "Subjective assessment methodology for preference of experience in 3dtv," in *IVMSP Workshop, 2013 IEEE 11th*, pp. 1–4, IEEE, 2013.

[23] ITU-T P.915, "Subjective assessment methods for 3d video quality," *International Telecommunication Union* , Mar. 2016.

[24] IEEE P3333.1.1, "Standard for the quality of experience (qoe) and visual comfort assessments of three dimensional (3d) contents based on psychophysical studies," 2015.

[25] IVY Lab stereoscopic video dataset *Available: http://ivylab.kaist.ac.kr/demo/ivy3D-LocalMotion/index.htm* .

[26] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Depth estimation reference software (ders) 5.0," *ISO/IEC JTC1/SC29/WG11 M* **16923**, 2009.

[27] R. Bradley, "14 paired comparisons: Some basic procedures and examples," *Handbook of Statistics* **4**, pp. 299–326, 1984.

[28] R. Bradley and M. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika* **39**, pp. 324–345, Dec. 1952.

[29] W. Chen, J. Fournier, M. Barkowsky, P. Le Callet, *et al.*, "New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone," *Proceeding of Stereoscopic Displays and Applications XXII, SPIE 2011* , 2011.