Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three di erent designs

Jing Li, Marcus Barkowsky, Patrick Le Callet

▶ **To cite this version:**

Jing Li, Marcus Barkowsky, Patrick Le Callet. Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three di erent designs. SPIE Electronic Imaging, Stereoscopic Displays and Applications, Human Factors 2013, Feb 2013, San Francisco, United States. pp.1-12, 10.1117/12.2002075 . hal-00789033

HAL Id: hal-00789033
https://hal.science/hal-00789033

Submitted on 15 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs

Jing Li, Marcus Barkowsky and Patrick Le Callet

LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597
Polytech Nantes, rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France

ABSTRACT

The pair comparison method is often recommended in subjective experiments because of the reliability of the obtained results. However, a drawback of this method is that the number of comparisons increases exponentially with the number of stimuli, which limits its usability for a large number of stimuli. Several design methods that aim to reduce the number of comparisons were proposed in the literature. However, their performances in the context of 3DTV should be evaluated carefully due to the fact that the results obtained from a paired comparison experiment in 3DTV may be influenced by two important factors. One is the observation error from observer's attentiveness, in particular inverting the vote. The second factor concerns the dependence on the context in which the evaluation takes place. In this study, three design methods, namely Full Paired Comparison method (FPC), Square Design method (SD) and the Adaptive Square Design method (ASD) were evaluated by subjective visual discomfort experiment in 3DTV. The results from the FPC method were considered as the ground truth. Comparing with the ground truth, the ASD method provided the most accurate results with a given number of trials. It also showed the highest robustness against observation errors and interdependence of comparisons. Due to the efficiency of the ASD method, paired comparison experiments become feasible with a reasonably large number of stimuli for measuring 3DTV visual discomfort.

Keywords: Pair comparison, Visual discomfort, 3DTV, Adaptive Square Design, Subjective experiment

1. INTRODUCTION

Visual discomfort induced by watching 3D images or videos is getting more and more attention recently [1][2][3][4] [5][6][7] as it decreases severely the Quality of Experience (QoE) of the viewers . Consequently, the measurement of the degree of visual discomfort became an important topic recently.

Due to the added value of depth in 3D image/video, subjective assessment on visual discomfort in 3DTV is more complex than the traditional 2D video quality assessment. Viewers are not used to 3D television and not familiar with the concept of "visual discomfort", thus he/she has no reference to compare with as in the 2D condition. It might be difficult for the viewers to assign an absolute psychophysical scale to the stimulus. Pair comparison is considered as a more reliable method because firstly, it is easy for observers since they only need to provide their preference on each pair. Secondly, pair comparison is more sensitive to the conditions with small differences, thus, it improves the discriminability of the votes for the stimuli[8]. Therefore, pair comparison is widely adopted in recent studies of 3DTV. For example, in the literatures of [5][6] [8] and [9], the authors used the pair comparison methods to evaluate the visual discomfort, QoE and stereoscopic image quality.

Though pair comparison shows its advantages in subjective tests of 3DTV, it has a big practical issue in real applications, that is: with an increasing number of stimuli, the number of comparisons increases exponentially. In our previous paper[10], we analyzed a balanced sub-set pair comparison method called "Square Design" which was proposed by Dykstra[11]. Dykstra defined a formula to calculate "efficiency" to evaluate this method which showed that this method was highly efficient in predicting the scores of the stimuli. However, this "high efficiency" was validated under the condition of some assumptions, e.g., no observation errors (no observer voting

Further author information: (Send correspondence to Jing Li)
Jing Li: E-mail: jing.li2@univ-nantes.fr, Telephone: +33 (0)6 69 05 95 53

erroneously during the test), no influence from the interdependency of other stimuli, etc. In our previous study we found that the original Square Design method was quite sensitive to the typical observation errors, which were induced by the number of observations and the observers' attentiveness to err on the voting. Thus, this method should be carefully considered before applying to the real subjective test. According to the analysis, an adaptive square design method was proposed in our previous study[10]. Its performance was validated by a Monte-Carlo simulation experiment which showed that it is more robust than the original Square Design method. In addition, it showed higher experimental efficiency than the Square Design method and the full Paired comparison method.

In this study, the performances of the three pair comparison design methods, namely, the full pair comparison method (FPC), the Square Design method (SD) and the adaptive square design method (ASD) are compared by the subjective experiments of visual discomfort in 3DTV. Due to the fact that the viewer's vote for pair comparison may be influenced by the observation errors or the interaction of the votes on stimuli, five subjective experiments were designed for comparison and analyzing. Experiment 1 was conducted by FPC method and used as the ground truth of the results. Experiment 2 and 3 are designed to compare SD and ASD methods under the influence of observation errors. Experiment 4 and 5 are designed for comparing SD and ASD method under the influence of irrelevant stimuli.

The rest of the paper is organized as follows. In Section 2, the three different designs for pair comparison methodology are introduced. Then, the Bradley-Terry model will be introduced in Section 3 which is the base for analyzing the pair comparison data. The five experiments conducted in this study are presented in Section 4 then follows the results in Section 5. Section 6 concludes the paper.

2. PAIR COMPARISON METHODOLOGIES

2.1 Full Pair Comparison method

Pair Comparison method has been suggested by ITU-T Recommendation P.910 [12] for evaluating the quality in multimedia services. The test stimuli (A, B, C, etc.) are generally combined in all the possible n(n - 1)/2 combinations AB, AC, BC, etc. If considering the displaying order, all the pairs of sequences should be displayed in both the possible orders (e.g. AB, BA), the number of combinations will raise to n(n - 1). After each pair a judgement is made on which element in a pair is preferred in the context of the test scenario.

To distinguish with other paired comparison designs, the PC abbreviation is replaced by FPC (Full Paired Comparison) in this paper.

2.2 Square Design method

Since it is unwieldy to run all pairs in paired comparison method, one possible way is to omit some pairs completely. Dykstra[11] proposed a "balanced sub-set" method, which means that for certain pairs (i, j) the comparison numbers n_{ij} is 0 while for all other pairs it is a constant $n_{ij} = n$. Each of the stimuli has the same frequency of occurrence in the whole experiment. Dykstra developed four types of balanced sub-set design: "Group divisible designs", "Triangular designs", "Square designs" and "Cyclic designs". The "Square design" (SD) is briefly introduced here.

Assuming the stimuli number $t = s^2$, the SD is constructed by placing the t stimuli into a square of size s. Only pairs which are in the same column or row are compared. For example, if there are $t = 9$ stimuli, stimulus 1, 2,...,t could be placed into a square matrix as following:

$$\begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{matrix}$$

In this design, only the pairs among stimuli $(1, 4, 7)$, $(2, 5, 8)$, $(3, 6, 9)$, $(1, 2, 3)$, $(4, 5, 6)$ and $(7, 8, 9)$ are compared. The total number of comparisons is $t(\sqrt{t} - 1)$.

In square design, when the stimulus number is 9, the paired comparison number is 9×(3-1)=18, compared to 9×8/2=36 for the complete method. As this method only runs part of the pairs, there must be a loss of information. Dykstra gave a definition called "efficiency" to evaluate this method, which showed that this

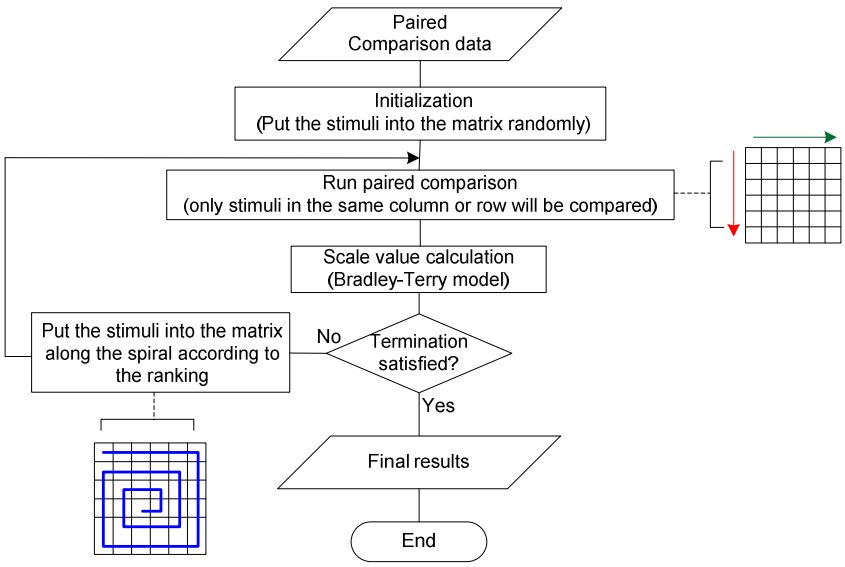


Figure 1. The flowchart of the ASD method for pair comparison.

method was highly efficient in predicting the scores of the stimuli. For details, the readers can be referred to [11].

2.3 Adaptive Square Design method

According to the Bradley-Terry model[13] which is a model to convert the paired comparison data to a scale value for each stimulus (more details can be found in Section 3), and the characteristics of the paired comparison[14], the SD method was examined in our previous paper[10] and we found that this method is not robust in the condition of observation errors. According to the analysis results, we pointed out that comparisons should be concentrated on the stimuli with similar quality. Thus, an adaptive Square Design was proposed in which the square matrix was updated after each observer's whole experiment. The detailed steps of this design are as following, and the flowchart of this method is shown in Fig. 1.

- Initialization of the square matrix. The position could be arranged randomly or according to the pre-test results. Afterwards, run paired comparisons.
- Calculation of the estimated scores. According to the current paired comparison results calculate the Bradley-Terry scores and sort them.
- Arrangement of the square matrix. According to the order rearrange the stimuli along the spiral, then run paired comparisons.
- Repeat step 2 and 3, until the termination conditions are satisfied (e.g., reach a certain number of comparisons or get the required confidence intervals).

The performance of the ASD method was evaluated by a Monte-Carlo simulation experiment in our previous study[10] but not the real subjective experiment. The simulation results in [10] showed that this method was robust to observation errors and more efficient than the SD method and the FPC method. The performance of this method in real application of 3DTV will be evaluated in the current study.

3. BRADLEY-TERRY MODEL

The Bradley-Terry model uses a linear model analyzing pair comparisons in order to map the probabilities of preference to scales. It introduces the term "merit" to describe the value of a particular stimulus on the scale. For the stimuli pair A_i and A_j, their "merits" are V_i and V_j respectively. This "merits" may represent a sensation magnitude on a scale, e.g., the degree of visual discomfort in 3DTV. During an observation, the observed "merit" values for A_i and A_j are X_i and X_j, the probability that the observer choose A_i over A_j is $P(X_i > X_j)$, which can be defined as:

$$P(X_i > X_j) \equiv \pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}, \qquad i \neq j \tag{1}$$

Where $\pi_i > 0$ and $\sum_{i=1}^{t} \pi_i = 1$, supposing there are in total t stimuli. The merit of stimulus A_i can be calculated by:

$$V_i = \log(\pi_i) \tag{2}$$

The Bradley-Terry score V_i is a negative value. Furthermore, the value V_i will change in function of the number of stimuli, since the sum of the π_i should equal to 1. However, for a certain pair A_i and A_j, their distance $V_i - V_j = \log \pi_i - \log \pi_j$ will not be changed since the π_{ij} is a constant. This characteristic of the Bradley-Terry score can be utilized to analyze the influence of observation errors and interdependence of comparisons on the final Bradley-Terry scores. In addition, the robustness of the paired comparison designs can be evaluated. More details can be found in Section 5.

Besides the Bradley-Terry score or sensation magnitude on a scale, the Bradley-Terry model can also provide confidence intervals, goodness of model fit and a series of hypothesis test. For more details, the reader is referred to [13][15].

4. EXPERIMENT

4.1 Apparatus

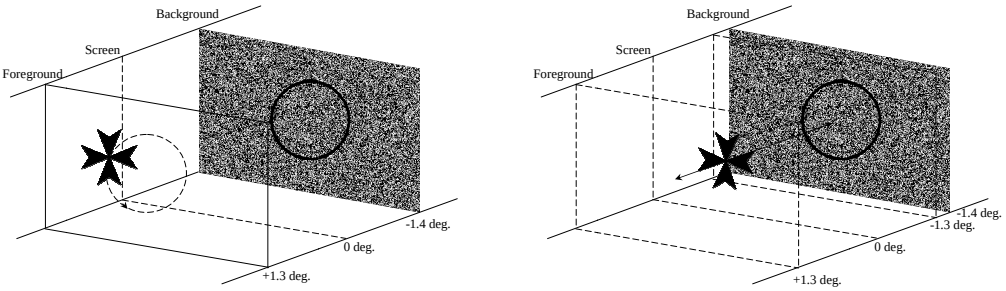
The display used in the experiment is a Dell Alienware AW2310 23-inch 3-D LCD screen (1920×1080 full HD resolution, 120Hz), which featured 0.265-mm dot pitch. The display was adjusted for a peak luminance of 50 cd/m^2 when viewed with the active shutter glasses. Stimuli were viewed binocularly through NVIDIA active shutter glasses (NVIDIA 3D vision kit) at a distance of about 90 cm. All environmental conditions were in line with ITU-R BT.500[16].

4.2 Stimuli

The stereoscopic sequences consist of a left-view and a right-view image which were generated by the MATLAB psychtoolbox [17]. A black Maltese cross was used as the foreground object with a size of 450×450 pixels (with the visual angle of 7.6 degree). The background was a salt and pepper-like noise image of 1920×1080 pixels. There are in total 36 stimuli in this study, including 15 planar motion stimuli, 5 static stimuli and 16 in-depth motion stimuli. The planar motion stimuli were exactly the same as our previous study [6], with 5 disparity levels (0, ± 0.65, ± 1.3 degree) and 3 velocity levels (slow, medium, fast). An example of the stimuli is shown in Fig. 2(a). For the static stimuli, the Maltese cross was positioned at the center of the screen, with five disparity levels which are 0, ± 0.65, ± 1.3 degree. For the in-depth motion stimuli, the Maltese cross was positioned in the center of the screen and moved forward and backward to the observer at different depth positions. An example is shown in Fig. 2(b).

4.3 Experimental design and assessment methods

The summary of the experiments is shown in Table 1. The details are illustrated by the following parts.



(a) (b)

Figure 2. Examples of the stimuli. (a) is an example of the planar motion stimuli. The foreground object moves at a depth plane with a disparity of 1.3 degree. The motion direction is anti-clockwise. The background is placed at the depth plane with a disparity of -1.4 degree. (b) is an example of the in-depth motion stimuli. The Maltese cross moves in depth between disparity +1.3 to -1.3 degree back and forth.

Table 1. Summary of the experiments

	Exp1	Exp2	Exp3	Exp4	Exp5
Assessment Method	FPC	SD	ASD	SD	ASD
Number of stimuli	15	16	16	36	36
Number of observers	45	33	33	33	33
Number of trials per observer	105	48	48	180	180

4.3.1 Experiment 1: FPC method for establishing ground truth

Fifteen planar motion stimuli were tested in the Experiment 1. The FPC method was used thus $15\times14/2 = 105$ pairs were presented in each individual subjective experiment. The results were used as the ground truth for the visual discomfort of the 15 planar motion stimuli.

In Experiment 1, the viewers watched a pair of stimuli at one trial, and then they were asked to select the one which made them more uncomfortable (the same for the Experiment 2 to Experiment 5). The presentation order for voting the whole set of 105 paired comparisons was randomly permuted for each viewer. The temporal presentation order of each pair of stimuli was balanced for all viewers.

4.3.2 Experiment 2 and 3: Comparing SD and ASD method under the influence of observation errors

The SD method was used in Experiment 2 and the ASD method was used in Experiment 3. Sixteen stimuli were tested in both experiments, including 15 planar motion stimuli and 1 static stimulus. The reason that we added one extra stimulus in this test is that sixteen is the required number to arrange all planar motion stimuli in a square matrix. We assumed that this extra stimulus would not generate significant influence on the final results. Thus, the main differences from this experiment and the Experiment 1 are the reduction of observations and the observation errors from observers which can be analyzed by comparing the test results and the ground truth.

The positions of the 16 stimuli in the square matrix were randomly assigned, as shown in the upper left 4×4 matrix in Fig.3. According to the SD and ASD methods, the stimuli in the same column or row will be compared which leads to 48 pairs for each observer. The only difference between SD and ASD method is that for the SD method, all observers watched the same pairs of stimuli. However, for ASD method, the initial positions of the stimuli are the same as in Experiment 2, but after the first observer's test, the positions of the stimuli will be updated for each observer according to all previous observers' results leading to different pairs compared by observers.

The presentation order for voting the whole set of 48 paired comparisons was randomly permuted for each viewer.

3	4	5	15	25	34
13	11	1	9	16	31
10	8	6	14	23	24
7	12	2	19	22	26
21	33	17	30	27	18
29	20	28	36	32	35

Figure 3. The arrangement of the stimuli in the square matrix for the SD method. The upper left 4×4 matrix is for Experiment 2 and 3. The whole matrix is for Experiment 4 and 5. Stimuli 1-15 represent the planar motion stimuli, stimuli 16-36 represent the other stimuli.

4.3.3 Experiment 4 and 5: Comparing SD and ASD method under the influence of irrelevant stimuli

The SD method was used in Experiment 4 and the ASD method was used in Experiment 5. Thirty-six video stimuli were tested in both experiments.

For Experiment 4, the upper left 4×4 matrix stays the same as in the Experiment 2. All the other positions were randomly placed by the remaining 20 stimuli as shown in Fig.3. In this way, the upper left 4×4 matrix can be considered as a copy of Experiment 2 except for the influence of the other stimuli from the remaining positions. Thus, the influence of the other stimuli on the results of the 15 planar motion stimuli can be analyzed by comparing the results of Experiment 2 and 3 and the test results from Experiment 4 and 5.

For Experiment 5, the initial positions of all stimuli were the same with the Experiment 4. After each observer's test, the positions of the stimuli will be updated according to the rule of ASD method.

According to the SD and ASD method, there are 180 pairs to be compared for each observer in both experiments. The presentation order for voting all 180 paired comparisons was randomly permuted for each viewer.

4.4 Observers

The number of observers for each experiment is shown in Table 1. It should be noted that the observers in Experiment 2 also participated in the Experiment 3. The observers are all non-experts in psychophysical studies on 3D, image processing or 3D related field. All have 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.

4.5 Procedures

The subjective experiment contained a training session and a test session. In the training session, there were five pairs of stimuli. At the beginning, the viewers were told that they will watch a series of synthetic stereoscopic motion images. They were asked not to stare at the moving object all the time, but watch the whole scene of the stereoscopic sequence under test. Then, they should select the one which made them more uncomfortable, concerning e.g., eye strain, headache. The viewers used two keys to switch between the pair of stimuli on a single screen. There was a minimum time limit for the display of each stimulus. Each observer had to watch the stimulus longer than the minimum time limit and then make a decision by pressing a specified button. For both planar motion and in-depth motion stimuli, the minimum time limit is defined as that the foreground moves along the circle one round or moves back and forth one round. For the static stimuli, the minimum time limit is 5 seconds. During the training session, all questions of the viewers were answered. We ensured that after the training session, all of the viewers understood the process and task of this experiment clearly.

For the main test session, the Experiment 1 contained 105 pairs, Experiment 2 and 3 contained 48 pairs, Experiment 4 and 5 contained 180 pairs. To avoid visual fatigue caused by long time watching affecting the experimental results, the Experiment 1, Experiment 4 and 5 were split into two sub-sessions. The viewers were asked to take a 10 minutes break after half of the test samples.

Table 2. The Correlation of the results with the Ground truth

Methods	CC	ROCC	RMSE
Exp2: SD - 4×4	0.9819	0.9536	0.2572
Exp3: ASD - 4×4	0.9913	0.9571	0.1623
Exp4: SD - 6×6	0.9590	0.9679	0.3261
Exp5: ASD - 6×6	0.9948	0.9857	0.1380

5. RESULTS

The Bradley-Terry model was used to analyze the subjective experiment results. The program used in this study for the Bradley-Terry model is available in [18].

Though the presented stimuli in Experiment 2 to Experiment 5 contained other types of stimuli, in this study, we only compare the results of the planar motion stimuli with the ground truth. Thus, the input of the Bradley-Terry model are the five Pair Comparison Matrices (PCM) with size of 15×15 for the planar motion stimuli. The output are the Bradley-Terry visual discomfort scores for the fifteen planar motion stimuli of each of the five experiments.

5.1 Comparative analysis

In this section, the general performances of the SD method and the ASD method will be compared through the correlations between the experimental results and the ground truth. According to our previous paper [6], the figures for planar motion stimuli are drawn according to the relative disparity between the foreground and the background, and the velocity. The results of all experiments are shown in Fig. 4.

Comparing the curves of the results from Experiment 2 to 5 with the ground truth, it can be found that the ASD method in Experiment 3 and 5 performed better than the SD method in Experiment 2 and 4. In particular, the curves from Experiment 4 matched the ground truth the worst. This result showed that the SD method may not generate reliable results under the condition of observation errors and the influence of the other stimuli.

To evaluate their correlations with the ground truth mathematically, the Pearson Linear Correlation Coefficient (CC), Spearman Rank-order Correlation Coefficient (ROCC) and the Root Mean Square Error (RMSE) are used as the criterions. The RMSE was calculated after linear mapping the tested data to the ground truth data because the Bradley-Terry score is not an absolute value. The results are shown in Table 2.

According to the CC, ROCC and RMSE values, the results indicated that when there are only observation errors from the observers, i.e., Experiment 2 and 3, the performance of the SD method is slightly worse than the ASD method but still comparable. However, when there are both observation errors and the influence from the existence of other stimuli, the SD method became less reliable. On the contrary, the ASD method remained robust in this conditions indicating its efficiency and reliability.

5.2 Quantitative analysis

In this section, we utilize the Bradley-Terry difference score matrix D to analyze the influence from observation error and from the presence of other stimuli. $D(i,j) = V_i - V_j$. V_i is the Bradley-Terry score of stimulus i. Thus, in this study, D is a 15×15 matrix. We use D_{gt}, $D_{sd4\times4}$, $D_{asd4\times4}$, $D_{sd6\times6}$ and $D_{asd6\times6}$ to represent the D matrices from Experiment 1 to Experiment 5, respectively.

5.2.1 Influence from observation errors

Observation errors come from two aspects: one is from observers' attentiveness, the other is from the reduced number of observations. Firstly the influence from observers' attentiveness is analyzed. According to the Bradley-Terry model, the distance between the two stimuli $V_i - V_j$ is related to P_{ij}, where

$$V_i - V_j = \log \frac{P_{ij}}{1 - P_{ij}} \tag{3}$$

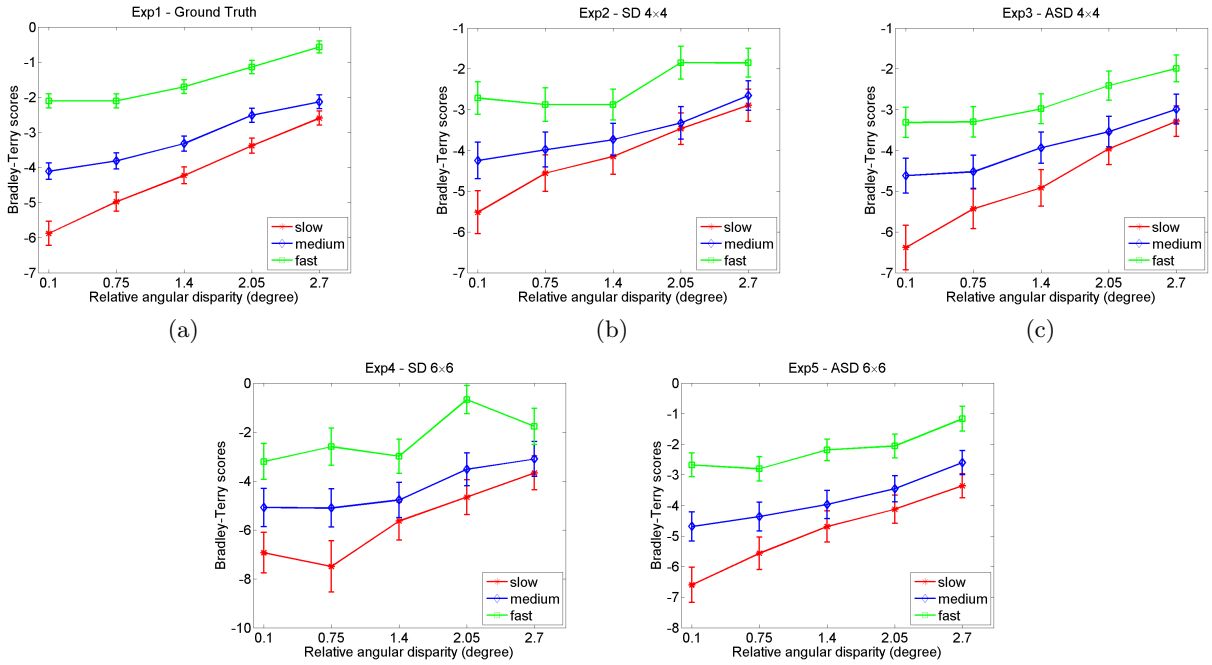


Figure 4. The Bradley-Terry scores for the planar motion stimuli. X-axis represents relative disparity, y-axis represents Bradley-Terry scores. Different lines in the figures represent different velocity levels, where slow, medium and fast represent 71.8, 179.5 and 287.2 degree/s. (a) is the ground truth data obtained by FPC method in Experiment 1. (b) is the experimental results of 4×4 SD method under the influence of observation errors (Experiment 2). (c) is the results of 4×4 ASD method under the influence of observation errors (Experiment 3). (d) is the results of 6×6 SD method under the influence of both observation errors and other stimuli (Experiment 4). (e) is the results of 6×6 ASD method under the influence of both observation errors and other stimuli (Experiment 5).

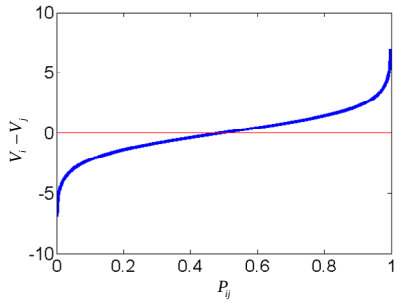


Figure 5. The relationship between the distance of two stimuli and the corresponding P_{ij} value. X-axis represents P_{ij}, y-axis represents the distance $V_i - V_j$.

The relationship between P_{ij} and $V_i - V_j$ is also shown in Fig. 5. It can be found that if the observation errors from observers' attentiveness are same, i.e., same change on P_{ij}, the influence on $V_i - V_j$ for the distant pairs would be larger than for the closer pairs.

Then, the influence from the reduced number of observations is analyzed. If P_{ij} can be represented by a fraction $P_{ij} = \frac{n}{m}, m > 1$, then,

$$V_i - V_j = \log \frac{n}{m-n} \tag{4}$$

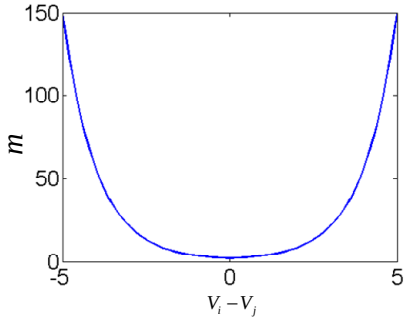


Figure 6. The relationship of the distance of two stimuli and the corresponding unit of number of comparisons. X-axis represents the distance $V_i - V_j$, y-axis m represents the minimum number of the comparisons to achieve this distance.

$$e^{V_i - V_j} = \frac{n}{m - n} \tag{5}$$

If assuming n=1, m represents the minimum number of comparisons that is required in order to achieve the distance $V_i - V_j$, then,

$$e^{V_i - V_j} = \frac{1}{m - 1}, m > 1 \tag{6}$$

The minimum number of comparisons m can be represented by the following equation:

$$m = \begin{cases} 1 + e^{V_i - V_j}, V_i \leq V_j \\ 1 + e^{-(V_i - V_j)}, V_i > V_j \end{cases} \tag{7}$$

This equation is depicted by Fig. 6. As shown in the figure, with the increase of the distance between two stimuli, the required comparison number is increasing exponentially. Thus, if the number of comparisons is reduced, the obtained p_{ij} values would generate more errors on estimating the $V_i - V_j$ value for distant pairs than for similar pairs.

To analyze the influence of observation errors on the experimental results, the difference between D_{gt} and $D_{sd4\times4}$, $D_{asd4\times4}$ is calculated, only the upper-right part of the matrix is considered, i.e.,

$$\begin{aligned} C_{gt,sd4\times4}(i,j) &= D_{gt}(i,j) - D_{sd4\times4}(i,j), i < j \\ C_{gt,asd4\times4}(i,j) &= D_{gt}(i,j) - D_{asd4\times4}(i,j), i < j \end{aligned} \tag{8}$$

The histogram of the errors C are shown in Fig. 7, with the corresponding fitted gaussian curve. μ, σ^2 represent mean and variance of the gaussian curve.

The observers in Experiment 2 and 3 were same, thus, we assumed that there is no influence from the observers on the two experiments. As shown in Fig. 7, the mean shift, the variance of the histogram for the SD method are larger than for the ASD method, which indicated that when the observation number is small and the raw pair comparison data are influenced by the observers' mistakes, the ASD method may be more reliable than the SD method.

5.2.2 Influence from the irrelevant stimuli

In this section, we analyze another factor that may affect the results which is the influence from irrelevant stimuli. In Experiment 4 and 5, besides the 15 planar motion stimuli, 21 further stimuli are added. 36 stimuli were arranged in a 6×6 matrix with the upper left sub-matrix exactly the same with the Experiment 2 and 3. Thus, in this test, both the observation errors and the influence of the added stimuli will affect the results. Assuming the observation errors can be eliminated by subtracting the results from Experiment 2 and 3, then,

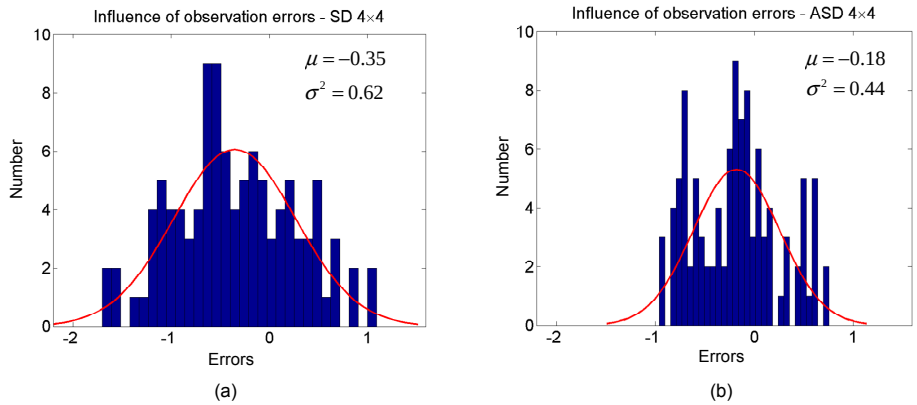


Figure 7. The histograms of $C_{gt,sd4\times4}$ and $C_{gt,asd4\times4}$. The red curves are fitted gaussian curve with mean values and variances. (a) is the results of Experiment 2. (b) is the results of Experiment 3.

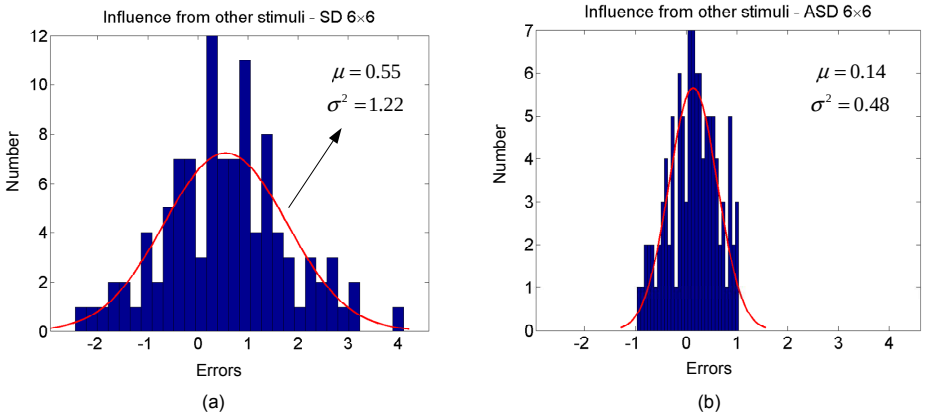


Figure 8. The histogram of $C_{sd4\times4-sd6\times6}$ and $C_{asd4\times4-sd6\times6}$. The red curves are fitted gaussian curve with mean values and variances. (a) is the results of Experiment 4. (b) is the results of Experiment 5.

the matrices $C_{sd4\times4-sd6\times6}$, $C_{asd4\times4-asd6\times6}$ represent the influence from other stimuli, as shown in the following Equation (9).

$$\begin{aligned} C_{sd4\times4-sd6\times6}(i,j) &= D_{sd4\times4}(i,j) - D_{sd6\times6}(i,j), i < j \\ C_{asd4\times4-asd6\times6}(i,j) &= D_{asd4\times4}(i,j) - D_{asd6\times6}(i,j), i < j \end{aligned} \tag{9}$$

The same processing as the previous section was performed, the histogram of the errors C are calculated and shown in Fig. 8, with the corresponding fitted gaussian curve. μ, σ^2 represent mean and variance of the gaussian curve.

This result indicates that the existence of other stimuli increased the uncertainty of the pair comparison results. As shown in the figure, the mean shift and the variance of the histogram of the SD method are larger than in the ASD method. The ASD method in this case still shows its robustness over the SD method.

5.3 Discussions

Generally speaking, the ASD method is more robust to both influences from observation errors and from the existence of other stimuli than the SD method.

When investigating the mean shift from the observation errors, both the SD and ASD method generated negative values, which means the distance between the stimuli pair of the ground truth in Experiment 1 is

smaller than that in Experiment 2 and 3 using SD and ASD method. One possible explanation is that due to the reduction of the observation number, for pairs with large distances, the reduced number of observations would lead to an overestimation of the distance. For example, for a pair with $P_{ij} = 0.9991$, the $D(i, j)$ would be 7, but if there is only limited number of comparisons, e.g., 33 observers, the obtained p_{ij} might be 1, which leads to an estimated $D(i, j)$ of infinite. Thus, the estimated distance became larger than the actual value.

For the Experiment 4 and 5, the mean shift from the influence of other stimuli is a positive value. One possible explanation may be that most of the added stimuli generated more visual discomfort than the planar motion stimuli, thus, the perceptual difference between the planar motion stimuli would be compressed which leads to the smaller distance than the ground truth.

6. CONCLUSIONS

This study focuses on the comparison of the performances of different paired comparison designs on visual discomfort subjective tests. The performance is provided regarding two aspects: 1) The accuracy of the methods; 2) The influence from observation errors and the irrelevant alternatives. The experimental results indicated that the ASD method provided more accurate results than the SD method given a certain number of trials and it also showed higher robustness against observation errors and dependence of comparisons. Due to the efficiency of the ASD method, paired comparison experiments become feasible with a reasonably large number of stimuli for measuring 3DTV visual discomfort.

ACKNOWLEDGMENTS

The participants of the subjective experiment are gratefully acknowledged. This work has been partly conducted within the scope of the JEDI (Just Explore Dimension) ITEA2 project which is supported by the French industry ministry through DGCIS and the PERSEE project which is financed by ANR (project reference: ANR-09-BLAN-0170).

REFERENCES

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

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

[3] Speranza, F., Tam, W., Renaud, R., and Hur, N., "Effect of disparity and motion on visual comfort of stereoscopic images," in [*Proceedings of SPIE*], **6055**, 60550B (2006).

[4] Lee, S., Jung, Y., Sohn, H., Ro, Y., and Park, H., "Visual discomfort induced by fast salient object motion in stereoscopic video," in [*Proceedings of SPIE*], **7863**, 786305 (2011).

[5] Li, J., Barkowsky, M., Wang, J., and Le Callet, P., "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)*], 1–8, IEEE (2011).

[6] Li, J., Barkowsky, M., and Le Callet, P., "The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos," in [*The third International Workshop on Quality of Multimedia Experience (QoMEX2011)*], 155–160, IEEE (2011).

[7] Lambooij, M., IJsselsteijn, W., and Heynderickx, I., "Visual discomfort of 3D TV: Assessment methods and modeling," *Displays* **32**(4), 209–218 (2011).

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

[9] Lee, J., De Simone, F., and Ebrahimi, T., "Subjective quality evaluation via paired comparison: application to scalable video coding," *Multimedia, IEEE Transactions on* **13**(5), 882–893 (2011).

[10] Li, J., Barkowsky, M., and Le Callet, P., "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment," in [*2012 IEEE International Conference on Image Processing (ICIP 2012)*], 1–4 (2012).

[11] Dykstra, O., "Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs," *Biometrics* **16**(2), 176–188 (1960).

[12] ITU-T., "P.910, subjective video quality assessment methods for multimedia applications," *International Telecommunication Union* (1999).

[13] Bradley, R., "14 Paired comparisons: Some basic procedures and examples," *Handbook of statistics* **4**, 299–326 (1984).

[14] Silverstein, D. A. and Farrell, J. E., "Quantifying Perceptual Image Quality.," in [*PICS'98*], 242–246 (1998).

[15] Bradley, R. and Terry, M., "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika* **39**(3/4), 324–345 (1952).

[16] ITU-R, "500-11, methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union, Geneva, Switzerland* (2002).

[17] Pelli, D., "The videotoolbox software for visual psychophysics: Transforming numbers into movies," *Spatial vision* **10**(4), 437–442 (1997).

[18] Wickelmaier, F. and Schmid, C., "A matlab function to estimate choice model parameters from paired-comparison data," *Behavior Research Methods* **36**(1), 29–40 (2004).