SUBJECTIVE ASSESSMENT METHODOLOGY FOR PREFERENCE OF EXPERIENCE IN 3DTV

Jing Li, Marcus Barkowsky, Patrick Le Callet

LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597
Polytech Nantes, rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France
{jing.li2,marcus.barkowsky,patrick.lecallet}@univ-nantes.fr

ABSTRACT

The measurement of the Quality of Experience (QoE) in 3DTV recently became an important research topic as it relates to the development of the 3D industry. Pair comparison is a reliable method as it is easier for the observers to provide their preference on a pair rather than give an absolute scale value to a stimulus. The QoE measured by pair comparison is thus called "Preference of Experience (PoE)". In this paper, we introduce some efficient designs for pair comparison which can reduce the number of comparisons. The constraints of the presentation order of the stimuli in pair comparison test are listed. Finally, some analysis methods for pair comparison data are provided accompanied with some examples from the studies of the measurement of PoE.

***Index Terms*—** Paired comparison, efficient methods, 3DTV, Quality of Experience, Preference of Experience

1. INTRODUCTION

With the success of the 3D technology applied in cinema entertainment, 3D technology for home entertainment is on the way. Studies on the improvement of the Quality of Experience (QoE) in 3DTV are thus getting more and more attention recently. Different from the traditional 2D condition, QoE in 3DTV is multi-dimensional which includes image perceptual quality, depth quantity and visual comfort [1]. In the Qualinet White paper published in 2012 [2], QoE is defined as "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the users personality and current state."

How to evaluate the QoE subjectively is still an open question nowadays. Some international groups, for example, the Video Quality Experts Group (VQEG) 3DTV project and the ITU-T Study Group 9 are working towards the standardization of stereoscopic quality assessment. In 2012, a recommendation for subjective methods for the assessment of stereoscopic 3DTV systems is delivered (ITU-R BT.2021)[3].

Researchers in 3DTV usually use the traditional 2D subjective assessment methods to measure QoE, e.g., DSIS, DSCQS, SSCQE suggested by ITU-R BT.500[4]. However, these methods were designed for 2D condition in which "Quality" is a one-dimension scale. Whether these methods are applicable in a multi-dimension scale is unknown. Pair comparison is thus considered as a more reliable method as observers just need to answer one question: "which one of the two 3D sequences do you prefer?". The QoE issue is then converted to "Preference of Experience" (PoE) issue.

PoE is used to specify the outcome of the subjective test using pair comparison. It represents the preference of the QoE of the observers as the observers provided their preferences between each two videos rather than an absolute scale value for each video sequence. The target of this paper is to introduce some assessment methods for PoE and the corresponding statistical analysis methods.

This paper is organized as follows. In Section 2, the standard methodology for PoE is introduced. Section 3 introduces some efficient pair comparison method. The constraints of the presentation order of the stimuli in pair comparison tests are listed in Section 4. Some analysis methods for the pair comparison are given in Section 5. Section 6 concludes the paper.

2. STANDARD PAIR COMPARISON METHODS AND RELATED MODELS

2.1. Full pair comparison method (FPC)

Full pair comparison method is already a standardized subjective video quality assessment method for multimedia applications [5]. For m stimuli S_1, S_2, ..., S_m, the test pairs are generated by combining all the possible $N = m(m-1)/2$ combinations $\{S_1S_2\}$, $\{S_1S_3\}$, $\{S_2S_3\}$, etc. If considering the displaying order, all the pairs of sequences should be displayed in both possible presentation orders (e.g. $\{S_1S_2\}$, $\{S_2S_1\}$), the number of combinations will raise to $N = m(m-1)$ for one observer. After the presentation of each pair a judgement is made on which element in a pair is preferred in the context of the test scenario.

The outcome of the FPC method is a pair comparison matrix **A**, where **A** = $(a_{ij})_{m\times m}$. a_{ij} is the total count of preference of stimulus S_i over S_j for all observers. $a_{ii} = 0$ for $i = 1, 2, ..., m$. The total number of comparisons for stimuli pair $\{S_iS_j\}$ is $n_{ij} = a_{ij} + a_{ji}$.

2.2. Pair comparison models

The Bradley-Terry (BT) model [6] and the Thurstone-Mosteller (T-M) model [7] are two well-known models to convert pair comparison data to psychophysical scale values for all stimuli. Basically, the pair comparison model is a function f, where

$$V_i - V_j = f(P_{ij}) \tag{1}$$

P_{ij} represents the probability that stimulus S_i is preferred to S_j, i.e., $P_{ij} = a_{ij}/n_{ij}$. The outputs are the differences of the scale values

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).

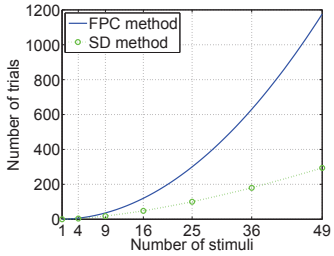


Fig. 1. Comparison of the trial numbers for FPC and SD method.

between stimuli S_i and S_j. By utilizing the least squares estimation or the maximum likelihood estimation, the scale value V_i for each stimulus, $i = 1, ..., m$ can be estimated.

3. EFFICIENT PAIR COMPARISON METHODS

FPC is a reliable method, however, the drawback of this method is that with the increase of the number of the stimuli, the number of comparisons increases exponentially which makes the subjective test infeasible. In this section, some efficient pair comparison methods which can reduce the number of comparisons are introduced. Most of these methods have been verified in our previous subjective visual discomfort tests on 3DTV, for details readers can refer to [8][9].

3.1. Square design (SD)

The SD method was proposed by Dykstra in 1960[10]. If the number of the stimuli m is a squared number $m = t^2$, the SD method is constructed by placing the indices of the m stimuli randomly into a square matrix $\mathbf{R} = (r_{pq})_{t\times t}$, where r_{pq} is the index of the Stimulus in position (p, q). Then the stimulus pair $\{S_iS_j\}$ are compared if and only if $(i, j) \in$ set $\boldsymbol{C}$, where $\boldsymbol{C}$ is defined as

$$\boldsymbol{C} = \{(x, y)|p = p' \vee q = q', \text{where } x = r_{pq}, y = r_{p'q'} \text{ in } \mathbf{R}\}$$

Thus, in this design, the number of comparisons for one observer is $N = m(\sqrt{m} - 1)$ compared to $m(m - 1)/2$ for the full pair comparison, which reduces the number of comparison significantly as shown in Fig.1. In addition, each stimulus has the same frequency of occurrence which makes a balance between the presence of the stimuli, thus it's called "balanced sub-set" design[10].

Here we give an example for SD method. Supposing $m = 9$, the matrix $\mathbf{R}$ may be designed as follows:

$$\mathbf{R} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

In this design, $\boldsymbol{C}$ = {(1,4), (1,7), (4,7), (1,2), (1,3), (2,3), (2,5)...}. Thus, only stimulus pair $\{S_1S_4\}$, $\{S_1S_7\}$, ... are compared. Each of the stimuli has the same frequency of occurrence which is 4 in this example.

Though the SD method showed "high efficiency" according to the analysis provided by Dykstra, this method is not robust to observation errors and the influence from the occurrence of other stimuli if the indices of the stimuli were placed into $\mathbf{R}$ randomly [8][9]. According to the analysis in [8], comparisons should be concentrated on the pairs with closer V_i (e.g., quality in quality assessment test) in the test. Thus, the stimuli pairs with similar V_i should be arranged in the same column or row to increase the probability to be compared.

There are several ways to implement this requirement. Two methods are proposed by the authors and will be introduced in Section 3.2 and 3.3.

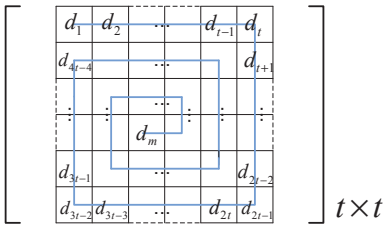


Fig. 2. The design for square matrix $\mathbf{R}_{OSD}$.

3.2. Optimized square design (OSD)

The "Optimized square design" is used for the conditions that the ranking of the stimuli in the test is known based on pre-test results or prior knowledge. Supposing the ordering indices of the stimuli (descending or ascending) is $\boldsymbol{d} = (d_1, d_2, ..., d_m)$. The square matrix is arranged in such a way that the elements of the vector $\boldsymbol{d}$ are placed along a spiral as shown in Fig. 2, which is defined as matrix $\mathbf{R}_{OSD}$.

Following the SD rule, the stimulus pair $\{S_iS_j\}$ is compared if and only if $(i, j) \in$ set $\boldsymbol{C}'$, where $\boldsymbol{C}'$ is defined as:

$$\boldsymbol{C}' = \{(x, y)|p = p' \vee q = q', \text{where } x = r_{pq}, y = r_{p'q'} \text{ in } \mathbf{R}_{OSD}\}$$

The square matrix $\mathbf{R}_{OSD}$ doesn't change for all observers.

3.3. Adaptive square design (ASD)

The "Adaptive Square Design" is proposed in the way that the square matrix $\mathbf{R}_{OSD}$ is updated for each observer. This method is used for the conditions that previous estimates are not available. The detailed steps of this design are as follows:

1. For the $1st$ observer, the square matrix used is $\mathbf{R}$ as introduced in Section 3.1, i.e., indices of the stimuli are randomly placed in $\mathbf{R}$. Run pair comparison experiment, only the pairs whose indices $\in \boldsymbol{C}$ are compared.
2. For the k_{th} observer ($k \geq 2$), according to the pair comparison matrix $\mathbf{A}$ of all previous $k - 1$ observers, the B-T scores and the ordering indices of the stimuli (descending or ascending) $\boldsymbol{d}^{k-1} = (d_1^{k-1}, d_2^{k-1}, ..., d_m^{k-1})$ are obtained ($\boldsymbol{d}^{k-1}$ represents the ordering indices vector after observer $k - 1$ finishing test). Based on vector $\boldsymbol{d}^{k-1}$, the square matrix $\mathbf{R}_{OSD}^k$ and $\boldsymbol{C}'^k$ are constructed as introduced in Section 3.2 ($\mathbf{R}_{OSD}^k$ and $\boldsymbol{C}'^k$ represents $\mathbf{R}_{OSD}$ and $\boldsymbol{C}'$ for the k_{th} observer). Run pair comparison experiment, only the pairs whose indices $\in \boldsymbol{C}'^k$ are compared.
3. Repeat step 2, until termination conditions are satisfied (e.g., all observers finished the test or targeted accuracy on confidence intervals are obtained).

3.4. Other designs

Besides the SD method, there are some other "balanced sub-set" designs, for example, "Group Divisible designs", "Triangular Designs" and "Cyclic Designs". For the condition that m is not a squared number, the SD method can be replaced by these designs. For more details please refer to [10]. The arrangement of the matrix of these designs should also follow the OSD or ASD rules, i.e., closer pairs should have higher probability to be compared, according to the test scenarios.

4. CONSTRAINTS ON THE PRESENTATION ORDER OF THE STIMULI

In the process of the stimulus presentation, an imbalance of the randomization of the stimuli would affect the paired comparison results significantly. Thus, the constraints on the stimulus randomization are defined:

- The presentation of the sequence content should be as random as possible, no observer watches the same content in two consecutive presentations.
- For each observer, the presentation order for each sequence should be balanced, i.e., $\{S_A S_X\}$, $\{S_Y S_A\}$. This means for all the pairs which include sequence S_A, half of the pairs should show S_A firstly, the rest should show S_A secondly.
- For all observers, all the pairs of sequences must be displayed in both orders. This means that the sequences that were displayed firstly are now displayed secondly and vice versa. For example, if one observer watches $\{S_A S_B\}$, there must be another observer watches $\{S_B S_A\}$.

5. DATA ANALYSIS METHODS

In this section, some analysis methods for PoE are introduced. As visual discomfort is an important dimension in QoE of 3DTV, in this section, our previous experimental results on visual discomfort [11] are used as examples to illustrate how to use the provided methods to analyze the data.

For the sake of naming the examples, the test conditions are introduced briefly. There are in total 15 stereoscopic synthetic video stimuli. Each stimulus consists of a planar moving foreground object and a fixed background image. The foreground has 3 velocity levels (slow, medium, fast) and 5 disparity levels (0, $\pm$ 0.65, $\pm$ 1.3 degree). In the following part, the stimulus is represented by (velocity, disparity). For more details the reader can be referred to [11]. 21 male and 24 female naive observers participated the test. The tests were conducted by FPC method.

5.1. Analysis on pair comparison raw data

5.1.1. Barnard's exact test

Barnard's test is a statistical significance test of the null hypothesis of independence of rows and columns in a contingency table, and it was claimed that it is more powerful than Fisher's exact test for 2×2 contingency tables [12]. Thus, in pair comparison data analysis, it may be used to check whether the P_{ij} is statistically significantly different from a probability of 0.5 (i.e., whether the observers are undecided), or whether there is significant difference between the P_{ij} of two conditions. For example, for a pair $\{S_i S_j\}$, in one test scenario, a_{ij} out of n_{ij} observers chose S_i. In another scenario, a'_{ij} out of n'_{ij} observers chose S_j. To compare if the test scenario would have influence on the comparison results, e.g., the test environment, the Barnard's exact test can be used here. The input of the test is a matrix $\begin{bmatrix} a_{ij} & a'_{ij} \\ n_{ij}-a_{ij} & n'_{ij}-a'_{ij} \end{bmatrix}$, the output of the test is a p-value. On a 95% confidence, p-value < 0.05 means there is significant difference between the probabilities that observers chose S_i over S_j of the two test scenarios.

Taking the data from male and female observers of our previous study as examples, gender can be considered as an influence factor. For the pair {A(slow, -0.65), B(fast, 0)}, P_{AB} = 3/24 for females, and P'_{AB} = 12/21 for males. The Barnard's exact test result provides p < 0.05, which means there is significant difference for the preference of the pair {(slow, -0.65), (fast, 0)} under the influence of gender. Females feel significantly more uncomfortable for the stimulus (fast, 0) than the male observers. Using Barnard's exact test to evaluate if there is preference between A and B for male observers, p = 0.52 > 0.05, which means there is no preference between these two stimuli for male observers.

5.1.2. Monte Carlo simulation test

After the Barnard's exact test on all pairs of the two conditions, we can obtain that in total a out of N pairs are significantly different. To test if the test conditions have influence on results, a Monte Carlo simulation experiment can be conducted by evaluating whether a/N is sufficiently large or not based on the observer's pair comparison data.

For example, in our previous experiment, there are in total 7 out of 105 pairs (7/105 = 0.0667) which are significantly different between male and female observers. To test if gender is an influence factor on the overall pairs, a Monte Carlo simulation experiment is conducted. This is based on comparing the ratio of the significantly different pairs in the test observer groups with the condition of randomly permuted two observer groups. The details are shown as follows:

Algorithm 1: Monte Carlo simulation experiment

Require: $Loop_num$ The number of loop;
N The number of stimulus pairs;
$\mathbf{A}_i$ The pair comparison matrix of observer i;
k = 0
while $k < Loop_num$ **do**
  $Group_1, Group_2 \leftarrow$
  Randomly divided all observers indices into two groups
  $(\mathbf{A}_{group1})_{m\times m} \leftarrow \sum_{i\in Group_1} \mathbf{A}_i$
  $(\mathbf{A}_{group2})_{m\times m} \leftarrow \sum_{j\in Group_2} \mathbf{A}_j$
  $e(k) \leftarrow$
  Number of significantly different pairs on $\mathbf{A}_{group1}, \mathbf{A}_{group2}$
  $Sig_ratio(k) \leftarrow e(k)/N$
  $k \leftarrow k+1$
end while
$\mu \leftarrow$ Mean of Sig_ratio
$\sigma \leftarrow$ Standard deviation of Sig_ratio
return μ, σ

If set $Loop_num$ sufficiently large (which also depends on the observer number), e.g., 1000, the distribution of the Sig_ratio can be estimated. The results are shown in Fig. 3. μ= 0.0822, σ = 0.04. Comparing with our test result 0.0667 which is lower than μ, we may conclude that the gender is not a main factor in the overall test though some of the pairs show significant difference.

5.2. Analysis based on pair comparison model

5.2.1. Bradley-Terry model and Thurstone-Mosteller model

The BT scores for the 15 planar motion stimuli are shown in Fig. 4(a). The scatter plot of the BT scores and the TM scores are shown in Fig. 4(b). The CC and ROCC for the output of the two models are 0.9997 and 0.9964, which verifies their similarity.

The BT model also provides the statistical analysis on whether there is significant difference between each two scores. The basic idea is to calculate the confidence intervals of the difference of the BT scores of each pair. If 0 is not in this interval, the scores of the

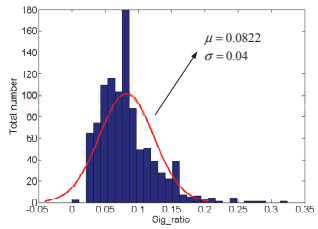
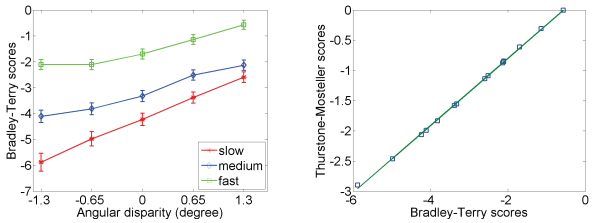


Fig. 3. Monte Carlo experiment results: The distribution of the Sig_ratio after 1000 times of simulation.



(a) (b)

Fig. 4. (a) Bradley-Terry scores for planar motion stimuli. X-axis represents disparity. Y-axis represents visual discomfort scores. Error bars represent 95% confidence intervals. (b) Scatter plot of the BT scores and the TM scores.

two stimuli are said to be significantly different. The calculation of the confidence interval can be found [13].

6. CONCLUSIONS AND DISCUSSIONS

In this paper, some efficient designs, namely, SD, OSD and ASD for pair comparison are introduced. SD method is not robust to observation errors in the real subjective test. The OSD and ASD method boost the pair comparison significantly though the time complexity is still far beyond the traditional SS or DSIS method. However, due to their robustness to observation errors, they are more reliable in measuring PoE of 3DTV.

Besides the traditional analysis methods, i.e., Bradley-Terry model, Thurstone-mosteller model, some innovative analysis methods, i.e., Barnard's exact test, Monte Carlo simulation test are introduced in this paper. Moreover, a model called "Elimination By Aspects (EBA)" model proposed in [14] are briefly introduced here for an open discussion.

According to EBA model, a subject prefers one stimulus over another due to a certain attribute that this stimulus has while the other does not. Stimuli without this attribute are eliminated from the set of possible alternatives. If all the stimuli under consideration share the preferred attribute, it will be disregarded for the current decision. Thus, another discriminating attribute has to be found, and the elimination process restarts [15]. The BT model in fact is a special case of the EBA model where there is only one attribute.

An application of the EBA model on measuring PoE in 3DTV might be in the case of an experiment to compare the influence of 2D and 3D technology on the PoE of the same video sequences. Thus, each video sequence has its own "quality" attribute. The presentation mode (2D or 3D) is another attribute which determines the observers' preference. The "quality" attribute for each video sequence and the "2D/3D" attribute for the presentation mode can be estimated by the EBA model. According to the comparison between the "2D" attribute and the "3D" attribute, which mode is preferred by the observers in a particular video content can be determined.

The EBA model provides a new perspective on analyzing the pair comparison data, which should be considered and employed in the future.

7. REFERENCES

[1] W. Chen, F. Jérôme, M. Barkowsky, and P. Le Callet, "Exploration of Quality of Experience of stereoscopic images: Binocular depth," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 1–6, Jan. 2012.

[2] P. Le Callet, S. Möller, and A. Perkis, "Qualinet white paper on definitions of quality of experience (2012)," Tech. Rep. version 1.1, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, June 2012.

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

[4] ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union, Geneva, Switzerland*, 2002.

[5] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union*, 1999.

[6] R.A. Bradley, "14 Paired comparisons: Some basic procedures and examples," *Handbook of statistics*, vol. 4, pp. 299–326, 1984.

[7] L.L. Thurstone, "A law of comparative judgment.," *Psychological review*, vol. 34, no. 4, pp. 273, 1927.

[8] J. Li, M. Barkowsky, and P. Le Callet, "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)*, 2012, pp. 1–4.

[9] J. Li, M. Barkowsky, and P. Le Callet, "Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs," *Proc. SPIE Electronic Imaging-Stereoscopic Displays and Applications XXIV*, 2013.

[10] O. Dykstra, "Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs," *Biometrics*, vol. 16, no. 2, pp. 176–188, 1960.

[11] J. Li, M. Barkowsky, and P. Le Callet, "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)*. 2011, pp. 155–160, IEEE.

[12] G. Barnard, "A new test for 2×2 tables," *Nature*, vol. 156:177, 1945.

[13] John C. Handley, "Comparative Analysis of Bradley-Terry and Thurstone-Mosteller Paired Comparison Models for Image Quality Assessment," in *Proc. IS&Ts Image Processing, Image Quality, Image Capture, Systems Conference*, 1081, pp. 108–112.

[14] A. Tversky and S. Sattath, "Preference trees," *Psychological Review*, vol. 86, no. 6, pp. 542–573, Nov. 1979.

[15] F. Wickelmaier and C. Schmid, "A matlab function to estimate choice model parameters from paired-comparison data," *Behavior Research Methods*, vol. 36, no. 1, pp. 29–40, 2004.