

A New Analysis Method for Paired Comparison and Its Application to 3D Quality Assessment*

Jong-Seok Lee
School of Integrated Technology
Yonsei University
406-840 Incheon, Korea
jong-seok.lee@yonsei.ac.kr

Lutz Goldmann, Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
{lutz.goldmann, touradj.ebrahimi}@epfl.ch

ABSTRACT

Among various subjective quality evaluation methodologies, paired comparison has the advantage of improved simplicity of the subjects' evaluation task due to simplified rating scales and direct comparison of two stimuli. Thus, it may lead to more reliable results when individual quality levels are difficult to define, quality differences between stimuli are small or multiple quality factors are involved. This paper proposes a new method to analyze results of paired comparison-based subjective tests. By assuming that ties convey information about significant differences between two stimuli being compared, the confidence intervals for the quality scores are estimated using a maximum likelihood criterion, which enables us to intuitively examine the significance of quality score differences. We describe the complete test methodology including the test procedure, outlier detection and score analysis applied to quality assessment of 3D images acquired using varying camera distances. Experimental results demonstrate the usefulness of the proposed analysis method, as well as the enhanced quality discriminability of the paired comparison methodology in comparison to the conventional single stimulus methodology.

Categories and Subject Descriptors

H.1.2 [Models and Principles]: User/Machine Systems—*Human Factors*; I.4.1 [Image Processing and Computer Vision]: Digitization and Image Capture—*Camera Calibration*

General Terms

Algorithms

Keywords

paired comparison, quality of experience, statistical significance, maximum likelihood estimation, stereoscopic image

*Area chair: Wei Tsang Ooi

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...\$10.00.

1. INTRODUCTION

Subjective multimedia quality assessment is to measure quality of experience (QoE) of multimedia content through subjective experiments. Its results provide knowledge about the way that human subjects perceive quality for the type of media considered. Such understanding can be exploited in various applications of multimedia, e.g. determination of content distribution strategies, development and benchmarking of multimedia processing algorithms, and development of objective quality metrics.

In general, it is important to utilize an appropriate test methodology for an experiment in order to exclude undesirable external factors, assess the targeted quality aspect accurately, and ensure reproducibility of the results. For this, there have been efforts to standardize test methodologies, e.g. [8]. Examples of frequently used methodologies are single stimulus and double stimulus methodologies. However, these methodologies have been developed mainly in the context of quality assessment for traditional 2D image/video content. Thus, applying them directly to assessment of new types of media such as 3D images/videos and high dynamic range images may not be as effective as for 2D image/video due to inherent differences in perception mechanisms. Often, quality differences between stimuli are too small, multiple quality factors are involved simultaneously, or individual quality levels are difficult to define, which may cause unreliable quality ratings by the subjects.

Paired comparison is one of the standardized test methodologies [8], which has potential to improve reliability of subjective tests due to the simplicity of the subjects' rating task. Instead of choosing a value on a discrete or continuous scale as in single stimulus or double stimulus methodologies, a subject only needs to indicate which one between a pair of stimuli has better quality. Results of paired comparison tests appear as winning frequencies between each pair and, thus, an additional analysis step is required to estimate a quality score and the corresponding confidence information for each stimulus.

In this paper, we propose a new method for analysing the ratings of the paired comparison in order to obtain quality scores equivalent to traditional mean opinion scores (MOS) and, more importantly, to the corresponding confidence intervals. The method is based on the assumption that ties carry information about significant differences of paired stimuli. Thereby, obtaining the confidence intervals is formulated as a maximum likelihood estimation problem in such a way that the intervals increase in proportion to the number of ties. The most important features of the proposed method

is that it enables intuitive examination of significant quality score differences.

To evaluate the proposed method, it is applied to the quality assessment of stereoscopic images acquired with varying inter-camera distances. Although 3D content is gaining popularity as a new type of media, many people still do not have much experience. Therefore, it is usually quite difficult for them to understand various quality factors involved in 3D perception and distinguish between quality variations in specific aspect under consideration. Through experiments, it is shown that the paired comparison methodology combined with the proposed analysis method can be successfully applied to 3D image quality evaluation and offers several advantages in comparison to the popular single stimulus methodology.

The rest of this paper is organized as follows. Section 2 provides a general overview of the paired comparison-based test methodology. Section 3 focuses on the novel analysis method to derive mean opinion scores and confidence intervals from the results of a paired comparison experiment. In Section 4, the effectiveness of the proposed method is demonstrated through a quality evaluation study for 3D images. Finally, concluding remarks are given in Section 5.

2. SIMULTANEOUS PAIRED COMPARISON

2.1 Test procedure

The basic idea of paired comparison is to present two stimuli simultaneously or sequentially and to ask subjects about their relative preferences of the stimuli. In general, a preference of stimulus A against stimulus B can be expressed on a continuous scale, e.g. $[-10, 10]$, or a discrete scale, e.g. {'much better', 'better', 'same', 'worse', and 'much worse'} [8]. In our case, however, we use a ternary scale, i.e. {'better', 'same', 'worse'}, in order to keep the subjects' rating task as simple as possible. Note that a weighting scheme can convert ratings done on a more subdivided scale to equivalent ratings on the ternary scale.

Inclusion of a tie (i.e. 'same') in the scale is useful for several reasons. First, it simplifies the subjects' rating task, especially when the quality difference between the given two stimuli is not easily noticeable. Moreover, if ties are not allowed, subjects are forced to choose between 'better' and 'worse' even when the quality difference of two stimuli appears unclear. This may result in biased results due to (possibly non-uniform) randomness. More importantly, ties provide useful information regarding the ambiguity in quality difference between two stimuli, which will be exploited for significance analysis in our method.

During the test session, each subject is asked to observe a pair of stimuli (A and B) and choose his/her preference between 'A is better', 'B is better', and 'same'. Especially, we employ simultaneous viewing by using two identical 3D displays, so that direct comparison between stimuli is possible and the reliability of ratings is enhanced.

2.2 Outlier detection

In single or double stimulus methodologies, the reliability of a subject is examined by comparing his/her ratings to those of other subjects. If the individual ratings deviate too much from the overall ratings, the subject is regarded as an outlier and his/her ratings are not considered further.

On the other hand, a subject's reliability in paired comparison can be evaluated by examining only his/her individual ratings across different pairs. Let $i > j$ indicate that stimulus i is rated better than j by a subject. If $i > j$, $j > k$ and $k > i$, a circular triad is formed, which violates the transitivity rule. If the number of circular triads is too large, the subject is regarded as an outlier. Since a tie (noted as $i = j$) is allowed in our rating scale, the following four cases are considered as circular triads: (1) $i > j$, $j > k$ and $k > i$, (2) $i > j$, $j > k$ and $k = i$, (3) $i > j$, $j = k$ and $k > i$, and (4) $i = j$, $j > k$, and $k > i$. If the ratio of the number of circular triads among all possible triads is larger than a threshold, the subject's ratings are discarded.

2.3 Quality score estimation

When ties are not allowed and the rating scale is binary, ratings by M subjects for a set of stimuli appear as winning frequencies, w_{ij} , $\forall i, j \in \{1, \dots, N\}$, representing the number of subjects who chose stimulus i against j . These frequencies need to be converted to quality scores equivalent to MOS for further quality examination and comparison.

A popular method for this is to use the Bradley-Terry (BT) model [1] that relates the empirical winning probability of stimulus i against j , $P_{ij} = w_{ij}/M$, with their quality scores, π_i and π_j , as:

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \quad (1)$$

The scores π_i ($i = 1, \dots, N$) satisfying $\pi_i \geq 0$ and $\sum_i \pi_i = 1$ can be obtained via maximum likelihood estimation.

When ties are included in the rating scale (as in our study), winning frequencies w_{ij} and tie frequencies $t_{ij} = t_{ji}$ summarize ratings of M subjects, where $w_{ij} + w_{ji} + t_{ij} = M$.

A few approaches have been proposed to accommodate ties in the framework of the BT model. A simple solution is to consider a tie as a half way between the other two preference options, i.e. $w_{ij} \leftarrow w_{ij} + t_{ij}/2$, and then to use the original BT model [3, 5]. The approaches proposed in [2, 6] introduced additional parameters that need to be estimated together with π_i .

3. PROPOSED ANALYSIS METHOD

Unlike the aforementioned approaches accommodating ties, the proposed method uses ties to infer the confidence of converted quality scores. Such confidence is important in analyzing and interpreting the quality scores. A higher quality score does not necessarily mean that the corresponding stimulus has a better quality than another stimulus if the score difference is not statistically significant. Therefore, confidence information enables us to prevent misleading judgment in such a case.

In our method, it is assumed that the ambiguity of quality differences between two stimuli is found in ties. The two extreme cases where the ties supposedly belong to one of the other two preference options are used to obtain the upper and lower bounds of the confidence intervals for the quality scores. Therefore, significance of quality differences between stimuli can be easily examined by checking whether their confidence intervals overlap. In other words, the fact that the confidence intervals of two stimuli overlap implies that the dominance of a score against the other may have been inverted and thus the quality difference is not significant.

First, the winning frequencies are used to obtain quality scores by using the BT model without considering ties. Then, the confidence interval for the score of each stimulus, $[\pi_i - \Delta\pi_i^-, \pi_i + \Delta\pi_i^+]$, is obtained as follows. The lower and upper bounds of the empirical winning probability of stimulus i , P_{ij}^- and P_{ij}^+ are obtained as

$$P_{ij}^- = w_{ij}/M \quad (2)$$

$$P_{ij}^+ = (w_{ij} + t_{ij})/M \quad (3)$$

assuming that the ties have been the preferences of stimulus j or i , respectively. These probabilities are related to the new parameters $\Delta\pi_i^-$ and $\Delta\pi_i^+$ as

$$P_{ij}^- = \frac{\pi_i - \Delta\pi_i^-}{\pi_i - \Delta\pi_i^- + \pi_j + \Delta\pi_j^+} \quad (4)$$

$$P_{ij}^+ = \frac{\pi_i + \Delta\pi_i^+}{\pi_i + \Delta\pi_i^+ + \pi_j - \Delta\pi_j^-} \quad (5)$$

A maximum likelihood criterion is used to estimate the values of $\Delta\pi_i^-$ and $\Delta\pi_i^+$ where the log-likelihood function to be maximized can be written as

$$L = \sum_{i=1}^N \sum_{j=1}^N \left\{ P_{ij}^- \log \left(\frac{\pi_i - \Delta\pi_i^-}{\pi_i - \Delta\pi_i^- + \pi_j + \Delta\pi_j^+} \right) + P_{ij}^+ \log \left(\frac{\pi_i + \Delta\pi_i^+}{\pi_i + \Delta\pi_i^+ + \pi_j - \Delta\pi_j^-} \right) \right\} \quad (6)$$

The aforementioned existing methods also reflect uncertainty in the resultant quality scores and the additional parameters in a way that their values change according to the number of ties. However, such relationship is less intuitive to examine significant quality score differences between stimuli in comparison to our method.

4. EXPERIMENT

4.1 Task and dataset

The EPFL 3D Image Database [4] was used in our study. It contains stereoscopic images with a resolution of 1920×1080 pixels covering diverse indoor and outdoor scenes with a large variety of colors, textures and depth structures. Each scene was captured using a stereo camera setup with six different inter-camera distances from 10 cm to 60 cm with a step size of 10 cm.

The camera distance is an important parameter in acquisition of stereoscopic images, which has significantly impact on the perceived 3D quality. A larger camera distance usually produces more 3D depth, but if it gets too large, it may cause unnaturalness and discomfort. The optimal camera distance depends on various factors including scene characteristics as well as the display size and viewing conditions.

In our experiments, six scenes were chosen from the database, i.e. sofa, tables, sculpture, moped, bikes, and construction. The other scenes were used for training. All possible pairs of each scene were considered to perform paired comparison tests. Thus, each subject watched $\binom{6}{2} \times 6 = 90$ pairs.

4.2 Environment

The tests were conducted at our laboratory designed for professional subjective quality tests according to the recommendations in [8]. The room walls were painted in gray 128

and the ambient lighting is achieved through a set of neon lamps with a color temperature of 6500 K. Two 46" LCD polarized stereoscopic displays with a native resolution of 1920×1080 pixels were used to present two stimuli simultaneously. Each subject sat in front of the two screens at a distance of approximately 2 m that is equivalent to 3 times the height of the screen. The subject was allowed to turn his/her head freely to watch the individual stimulus on each screen alternatively.

4.3 Procedure

Sixteen subjects (12 males and 4 females) participated in the tests. They were screened for visual acuity, color vision and binocular vision according to [7]. All of them were non-expert viewers with a marginal experience of 3D image and video viewing. Their ages ranged from 25 to 36 with an average of 29.9.

Prior to a test session, a training session was held to introduce the test procedure and rating task to the subject by using a set of training stimuli. In the middle of the test, a short break was given to prevent the fatigue of the subject.

The method described in Section 2.2 was used to detect outliers. Since the transitivity violation rates were less than 0.05 for all subjects, no subject was rejected as an outlier.

Quality scores and confidence intervals were obtained from the subjective ratings by using the method presented in Section 3. The scores were logarithmically scaled for better visualization, and then normalized so that the minimum and maximum scores for each content are 0 and 100, respectively.

4.4 Results and analysis

Figure 1 compares the results of paired comparison with those of single stimulus reported in [4]¹. In the latter case, the MOS and 95% confidence interval values are shown. For better comparison, the results of single stimulus tests were re-scaled to the same range as the paired comparison for each scene.

Overall, the results of the two test methodologies show a similar trend, i.e. the best quality for each scene is obtained for small camera distances that lead to a good 3D effect with comfortable disparity levels. As the camera distance increases, the quality decreases considerably due to the uncomfortable amount of disparity. Content-dependence of this trend is also observed, e.g. the construction scene shows relatively moderate quality decrease for camera distances of 20 cm to 40 cm due to the more distant background when compared to other scenes, which appears in both single stimulus and paired comparison results. For the case of paired comparison, the confidence intervals appear asymmetric around the quality scores because two separate variables were defined for the lower and upper bounds. This enables us to easily examine the significance of quality difference between two stimuli by checking the overlap between the confidence intervals of two stimuli.

An important advantage of paired comparison over the single stimulus methodology is that ambiguity between stimuli is decreased. In [4], it was mentioned that the subjects of the single stimulus tests had difficulty in discriminating quality differences especially for mid-range camera distances. However, such difficulty could be alleviated by simplifying the subjects' task through paired comparison,

¹Due to the page limit, the results for only three scenes are shown here.

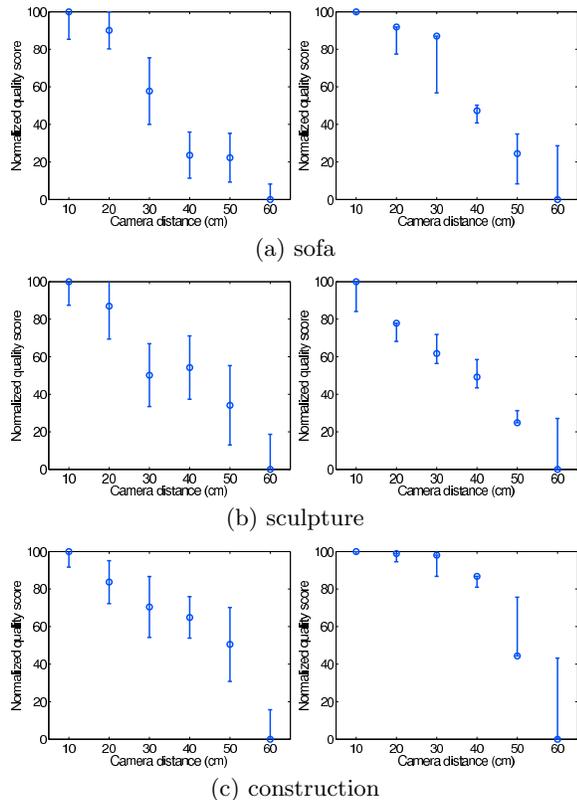


Figure 1: Comparison of the results of the single stimulus (left column) and paired comparison (right column) tests.

which resulted in enhanced discriminability between stimuli as it can be seen from the figure.

In order to compare the two methodologies in terms of discriminability, pair-wise comparison results were simulated from the quality scores of the single stimulus tests. In other words, the scores of a subject for each pair of stimuli were compared and, if the score difference is larger than a threshold, the one having a higher score was considered as the winner of the comparison, otherwise, a tie was recorded. By adjusting the threshold, we generated simulated comparison results having the same probabilities of ties to those of the paired comparison experiment. This procedure was repeated for all subjects involved in the single stimulus experiment. Then, we computed a discriminability measure defined as the absolute difference between the probabilities for the preference of each of two stimuli, i.e. $D_{ij} = |P_{ij} - P_{ji}|$. Table 1 compares the average values of the discriminability measure of the single stimulus and paired comparison methodologies. Overall, the paired comparison tests achieve higher discrimination between stimuli than the single stimulus tests.

5. CONCLUSIONS

We have proposed a new analysis method for paired comparison-based subjective quality assessment, which was applied to 3D image quality evaluation. The complete methodology including the test procedure, outlier detection, score calculation and significant analysis was described. Exploiting the ambiguity residing in ties, the proposed method facili-

Table 1: Average discriminability measures of single stimulus (SS) and paired comparison (PC) for each scene.

Method	sof.	tab.	scu.	mop.	bik.	con.	Avg.
SS	0.75	0.69	0.63	0.69	0.73	0.71	0.70
PC	0.73	0.72	0.75	0.84	0.79	0.83	0.78

tates intuitive examination of significant quality difference without any additional statistical hypothesis test. A case study on 3D image quality evaluation demonstrated that the paired comparison test methodology improves quality discriminability between stimuli in comparison to the single stimulus methodology.

It is worth mentioning that the improved discriminability was obtained at the expense of an increased test duration (i.e. 36 stimuli to be evaluated in a single stimulus test versus 90 stimulus pairs to be compared in a paired comparison test). In our future work, we will work on improving efficiency of paired comparison. We also plan to apply the proposed method to other evaluation tasks where conventional single or double stimulus methodologies have difficulty in obtaining reliable results.

6. ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Knowledge Economy, Korea, under the IT Consilience Creative Program (NIPA-2010-C1515-1001-0001), in part by Yonsei University Research Fund of 2011, in part by the European Network of Excellence PetaMedia (no. 216444), and in part by the COST Action IC1003 Qualinet.

7. REFERENCES

- [1] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.
- [2] R. R. Davidson. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Am. Stat. Assoc.*, 65(329):317–328, Mar. 1970.
- [3] M. E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *J. R. Stat. Soc., Ser. C (Appl. Stat.)*, 48(3):377–394, 1999.
- [4] L. Goldmann, F. D. Simone, and T. Ebrahimi. Impact of acquisition distortions on the quality of stereoscopic images. In *Proc. Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, Jan. 2010.
- [5] J.-S. Lee, F. D. Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, and T. Ebrahimi. Subjective evaluation of scalable video coding for content distribution. In *Proc. ACM Multimedia*, Firenze, Italy, Oct. 2010.
- [6] P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *J. Am. Stat. Assoc.*, 62(317):194–204, 1967.
- [7] Subjective assessment of stereoscopic television pictures. Recommendation ITU-R BT.1438, 2000.
- [8] Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-11, 2002.